ABSTRACT

HE, YUN. **Multiscale Signal Processing and Shape Analysis for an Inverse SAR Imaging System**. (Under the direction of **Prof. Hamid Krim**.)

The great challenge in signal processing is to devise computationally efficient and statistically optimal algorithms for estimating signals from noisy background and understanding their contents. This thesis treats the problem of multiscale signal processing and shape analysis for an Inverse Synthetic Aperture Radar (ISAR) imaging system. To address some of the limitations of conventional techniques in radar image processing, an information theoretic approach for target motion estimation is first proposed. A wavelet based multiscale method for shape enhancement is subsequently derived and followed by a regression technique for shape recognition.

Building on entropy-based divergence measures which have shown promising results in many areas of engineering and image processing, we introduce in this thesis a new generalized divergence measure, namely the *Jensen-Rényi* divergence. Upon establishing its properties such as convexity and its upper bound etc., we apply it to image registration for ISAR focusing as well as related problems in data fusion.

Attempting to extend current approaches to signal estimation in a wavelet framework, which have generally relied on the assumption of normally distributed perturbations, we propose a novel non-linear filtering technique, as a pre-processing step for the shapes obtained from an ISAR imaging system. The key idea is to project a noisy shape onto a wavelet domain and to suppress wavelet coefficients by a mask derived from curvature extrema in its scale space representation. For a piecewise smooth signal, it can be shown that filtering by this curvature mask is equivalent to preserving the signal pointwise Hölder exponents at the singular points, and to lifting its smoothness at all the remaining points.

To identify a shape independently of its registration information, we propose matching two configurations by regression, using notations of general shape spaces and procrustean distances. In particular, we study the generalized matching by estimating mean shapes in two dimensions. Simulation results show that matching by way of a mean shape is more robust than matching target shapes directly.

Report Documentation Page		*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE **JUN 2001**	2. REPORT TYPE	3. DATES COVERED **00-00-2001 to 00-00-2001**
4. TITLE AND SUBTITLE **Multiscale Signal Processing and Shape Analysis for an Inverse Sar Imaging System**		5a. CONTRACT NUMBER
		5b. GRANT NUMBER
		5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)		5d. PROJECT NUMBER
		5e. TASK NUMBER
		5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **North Carolina State University,Department of Electrical Engineering,Raleigh,NC,27695**		8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR'S ACRONYM(S)
		11. SPONSOR/MONITOR'S REPORT NUMBER(S)
12. DISTRIBUTION/AVAILABILITY STATEMENT **Approved for public release; distribution unlimited**		
13. SUPPLEMENTARY NOTES **The original document contains color images.**		
14. ABSTRACT		
15. SUBJECT TERMS		

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES **175**	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**			

MULTISCALE SIGNAL PROCESSING AND SHAPE ANALYSIS FOR AN INVERSE SAR IMAGING SYSTEM

by

YUN HE

A dissertation submitted to the Graduate Faculty of

North Carolina State University

in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

ELECTRICAL ENGINEERING

Raleigh

June, 2001

APPROVED BY:

Professor Hamid Krim	Professor Brian L. Hughes
Chair of Advisory Committee	

Professor Alexandra Duel-Hallen	Professor Siamak Khorram

To my wife and parents
for their endless love

Biography

Yun He received his B.S. and M.S. degrees in electrical engineering from Beijing University of Aero. & Astro., Beijing, China, in 1995 and 1998 respectively. From 1996 to 1998, he worked as a part-time design engineer of firmware development for networking applications. In May 1998, he started his Ph.D. study in electrical engineering at North Carolina State University, Raleigh, NC. His research interests are in communications and signal processing.

Acknowledgments

This endeavor was truly a learning experience. Thanks are due to many people for their interactions and collaborations.

I consider it a privilege to have worked under the supervision of Professor Hamid Krim, and owe him a great deal for his guidance, patience, and financial support through generous grants from Air Force Office of Scientific Research (AFOSR) and Multidisciplinary University Research Initiative (MURI). I have benefited immensely from his insight, wisdom, suggestions, comments, and constructive criticism of my work. I especially thank him for teaching me how to see the "big picture" as related to my research.

I would also like to thank my committee members Professor Brian L. Hughes and Professor Alexandra Duel-Hallen for their expertise and advice. I learned the fundamentals of communication and information theory through their classes and seminars. Their way of teaching showed me an excellent example of a systematic and straightforward fashion in presenting technical ideas. I am also grateful to have Professor Siamak Khorram on my thesis committee board. I thank him for his inquisitive comments and numerous suggestions in accuracy assessment, which led to many improvements in this thesis.

I also want to thank Dr. Victor Chen from Naval Research Laboratory for providing me the experimental data; and numerous developers of WaveLab, which is a Matlab toolbox freely available over the Internet. Without WaveLab, I guess I would have to spend at least one more year developing all the software.

It was a great pleasure to have closely worked with Dr. A. Ben Hamza and Dr. Yufang Bao. As coworkers, I thank them for the numerous suggestions, guidance, comments, and heated, friendly and enlightening discussions. I also want to thank my fellow graduate students for being good officemates, especially Oleg Poliannikov, for sharing coke and sense of humor which made many tough moments so brief.

I could not fail to thank our administrative assistant, Sandy Bronson, for her pa-

tience and kind help in numerous paperwork, travel arrangement and equipment orders. She is really a life saver.

Least but by far not last, I would like to thank my parents and my wife for their love, support and encouragement during my graduate studies. Mere words cannot express how much I love and appreciate them. This thesis is a dedication for their love.

Contents

x

List of Tables

List of Figures

1

Introduction

INVERSE Synthetic Aperture Radar (ISAR) is an imaging technique that achieves a high resolution by exploiting the relative motion between a stationary radar and a moving target. This is accomplished by coherently processing the returned radar signals so as to synthesize the effect of a larger aperture array laid out along the target's path of motion. One important application of ISAR is as a front-end system for the purpose of target recognition. The fundamental goal is to detect and recognize objects of interest in a noisy environment.

A typical ISAR imaging system consists of image acquisition, data fusion, target shape extraction and enhancement, and finally shape recognition. The main theme of this thesis is focused on information theoretic imaging and shape analysis. An information theoretic approach for ISAR image focusing and fusion is first proposed to serve as a robust stage for image acquisition. A wavelet based multiscale method for shape enhancement is subsequently derived to estimate a shape in a noisy background which is followed by a regression technique for shape recognition.

1.1 Problem Motivation and Formulation

Motivated by the current limitations of conventional techniques in radar image focusing, target shape estimation and recognition, this thesis addresses the following issues.

• **Information theoretic approach for ISAR image focusing**

The ISAR imagery is induced by the target rotation, which in turn causes time varying spectra of the reflected signals and blurs the resulting image. When a target exhibits complex motion, such as rotation and maneuvering, a standard motion compensation is not adequate to generate an acceptable image for viewing and analysis. In this thesis, we tackle this problem by an information theoretic approach.

In the work of Woods [1] and Viola [2], mutual information, a basic concept from information theory, is introduced as a measure of evaluating the similarity between images. When the two images are properly matched, corresponding areas overlap, and the resulting joint histogram contains high values for the pixel combinations of the corresponding regions. When the images are mis-registered, non-matched areas also overlap and will contribute to additional pixel combinations in the joint histogram. In case of misregistration, the joint histogram has less significant peaks and is more dispersed than that of the correct alignment of images. The registration criterion is hence to find a transformation such that the mutual information of the corresponding pixel pair intensity values in the matching images is maximized. This approach is accepted widely [3] as one of the most accurate and robust registration measures. Following the same argument, Hero [4] *et al.* extends the concept to apply Rényi entropy to measure the joint histogram as a dissimilarity metric between images.

Inspired by this previous work, and looking to address their limitation in often difficult imagery, we introduce in this paper a novel generalized information theoretic measure, namely the *Jensen-Rényi* divergence and defined in terms of Rényi entropy [5]. Jensen-Rényi divergence is defined as a similarity measurement among any finite number of weighted probability distributions. Shannon mutual information is a special case of the Jensen-Rényi divergence. This generalization provides us an ability to control the measurement sensitivity of the joint histogram.

The objective of ISAR image registration is to estimate the target motion during the imaging time. Let $\mathcal{T}_{(l,\theta,\gamma)}$ be a Euclidean transformation with translational parameter $l = (l_x, l_y)$, rotational parameter θ and scaling parameter γ. Given two ISAR image frames f_1 and f_2, the estimates of target motion parameters $(l^*, \theta^*, \gamma^*)$ are given by

$$(l^*, \theta^*, \gamma^*) = \arg \max_{(l,\theta,\gamma)} D_{JR\boldsymbol{\omega}_{\alpha}}(f_1, \mathcal{T}_{(l,\theta,\gamma)} f_2) \tag{1.1}$$

where $D_{JR\boldsymbol{\omega}_{\alpha}}(\cdot)$ is an induced similarity measure based on the proposed Jensen-Rényi divergence of order α and weight $\boldsymbol{\omega}$, which is maximal when f_1 matches $\mathcal{T}_{(l,\theta,\gamma)} f_2$. As the radar tracks the target, the reflected signal is continuously recorded during the imaging time. By registering a sequence of consecutive image frames, $\{f_i\}_{i=0}^{N}$, the target motion during the imaging time can be estimated by interpolating $\{(l_i, \theta_i, \gamma_i)\}_{i=1}^{N}$. Based on the estimated trajectory of the target, translational motion compensation (TMC), and rotational motion compensation (RMC) [6] can be used to generate a focused image of the target.

- **Multiscale Signal Enhancement: Beyond the Normality and Independence Assumption**

To fulfill the goal of an ISAR imaging system of recognizing objects of interest in a noisy environment, shape enhancement is required and constitutes a crucial step. Donoho and Johnstone [7] first showed that effective noise suppression may be achieved by wavelet shrinkage, in comparison to traditional linear methods. Given the noisy wavelet coefficients, i.e. the true wavelet coefficients plus a noise term, and assuming that one has knowledge of the true wavelet coefficients, an ideal filter sets a noisy coefficient to zero if the noise variance σ^2 is greater than the square of the true wavelet coefficient; otherwise the noisy coefficient is kept. In this way, the mean square error of this ideal estimator is the minimum of σ^2 and the square of the coefficient. Under the assumption of i.i.d. normal noise, it can shown that a soft thresholding estimator achieves a risk at most $O(\log M)$ times the risk of this ideal estimator, where M is the length of the observation.

To choose an appropriate threshold, Donoho and Johnstone [7] have taken a minimax

approach to characterize the signal, and they proved that the estimation risk is close to the minimax risk by setting a threshold

$$T = \sigma \sqrt{2 \log_e M}.$$

Krim and Pesquet [8] have given an alternative derivation for this threshold, using Rissanen's Minimum Description Length (MDL) criterion [9] and the assumption of normally distributed noise. The threshold T increase with M is due to the tail of the Gaussian distribution, which tends to generate larger noise coefficients when sample size increases. This threshold is not optimal, and in general a lower threshold reduces the risk. To refine the threshold, a SureShrink [10] procedure is proposed. Sureshrink calculates thresholds by the principle of minimizing the Stein unbiased estimate of risk for threshold estimates. SureShrink is also based on the assumption of i.i.d. normal noise, which does not hold for ISAR applications. For non-Gaussian type of noise, Neumann [11], Averkamp and Houdre [12] studied the choice of thresholds by having recourse to asymptotics. Wavelet thresholding theory is, however, based on the assumption that we know the statistics of the noise to determine an adequate threshold. This makes the algorithm less flexible and less adaptive to different scenarios which can result in an even worse reconstruction. Compensation for the lack of a prior knowledge of the noise statistics may be handled by adopting the minimax principle [13] upon deriving the worst case noise distribution.

Points of sharp variations are often among the most important features for analyzing properties of transient piecewise smooth signals. To characterize the singular structures, Hölder exponents [14] provide a pointwise measure of a function over a time interval. Due to the pioneering work by Jaffard [15] and Meyer [16], it can be shown that a local signal singularity of a function is characterized by the decay of its wavelet transform amplitude across scales.

In this thesis, a novel nonlinear filtering technique is proposed. We assume that a prior knowledge about pointwise smoothness measure of the signal is known or can be extracted. However, this smoothness property of the signal is corrupted by additive noise, which in general has a uniform Hölder exponent less than 1. We view the denoising

problem as one of carefully controlling the Hölder exponents of measured data with a goal of extracting the signal portion with some smoothness fidelity to the original. Let $f \in \mathbf{L}^2(\mathbb{R})$. We consider an additive noise model. The measured data are

$$Z(t) = f(t) + N(t), \tag{1.2}$$

where the noise is modeled by the realization of a zero mean random process N. Denote $\alpha_f(\cdot)$ and $\alpha_N(\cdot)$ characterize the pointwise Hölder exponent of $f(\cdot)$ and $N(\cdot)$ respectively. A ideal operator T satisfies the following two conditions:

$\rightarrow \hat{f} = TZ$ admits $\alpha_f(\cdot)$ as its pointwise Hölder exponent

$\rightarrow V(t) = Z(t) - \hat{f}(t)$ admits $\alpha_N(\cdot)$ as its pointwise Hölder exponent

This non-linear filter is optimal in the sense of recovering the smoothness of the true underlying signal.

• **Information theoretic approach for multiscale image fusion**

With the development of new imaging sensors arises the need for image processing techniques that can effectively fuse images from different sensors into a single coherent composition for interpretation. In order to make use of inherent redundancy and extended coverage of multiple sensors, we propose a multiscale approach for pixel level image fusion. The ultimate goal is to reduce human/machine error in detection and recognition of objects.

Over the past two decades, a wide variety of pixel-level image fusion algorithms has been developed. These techniques may be classified into linear superposition, logical filter [17], mathematical morphology [18], image algebra [19] [20], artificial neural network [21], and simulated annealing [22] methods. Each of these algorithms focuses on the fact that the fused image reveals new information concerning features that can not be perceived in individual sensor images. However, some useful information has been discarded since each fusion scheme tends to emphasize different attributes of the image. Luo [23] provides a detailed review of these techniques.

Inspired by the fact that the human visual system processes and analyzes image information at different scales, researchers recently proposed a multiscale based fusion

method which is widely accepted [24] as one of the most effective techniques for image fusion. Wavelet theory has played a particularly important role in multiscale analysis. A number of papers [25] [26] [27] have addressed fusion algorithms based on the orthogonal wavelet transform. A major drawback in the recent pursuit of wavelet-based fusion algorithms is due to a lack of a good fusion scheme. Most fusion rules so far proposed are in essence more or less similar to "choose max" scheme proposed by Burt [28], which introduces a significant amount of high frequency noise due to the sudden switch of the fused wavelet coefficient to that which is maximum of the source. This high frequency noise is particularly undesirable to visual perception.

In this thesis, we apply a biorthogonal wavelet transform to the pixel level image fusion. It is possible to construct smooth biorthogonal wavelets of compact support which are either symmetric or antisymmetric. At the exception of a Haar wavelet, it has been shown [29] that symmetric orthogonal wavelets are impossible to construct. Symmetric or antisymmetric wavelets are synthesized with perfect reconstruction filters having a linear phase. This is a desirable property for image fusion applications.

Unlike the "choose max" type of selection rules, we propose an information theoretic fusion scheme. For each pixel in a source image, a vector consisting of wavelet coefficients at that pixel position across scales is formed to indicate the "activity" of that pixel. We denote these indicator vectors of all the pixels in a source image as its activity map. To make a reasonable comparison among activity indicator vectors, we apply our newly proposed divergence measure, Jensen-Rényi divergence, which is defined in terms of Rényi entropy.

Let $f_1, f_2, ..., f_n : \mathbb{Z}^2 \to \mathbb{R}$ be digital images of the same scene taken from different sensors. Denote

$$W f_i = \{d_i^1(j, \boldsymbol{n}), d_i^2(j, \boldsymbol{n}), d_i^3(j, \boldsymbol{n}), a_i(L, \boldsymbol{n})\}_{0 < j \leq L, \boldsymbol{n} \in \mathbb{Z}^2}, \ i = 1, 2, ... n$$

as a biorthogonal wavelet image representation of f_i. Our fusion scheme cab be formulated as the following optimization problem:

$$W f = \arg \min_{W f \in \mathcal{F}} \left\{ \sum_{i=1}^{n} (\sum_{j=1}^{L} \sum_{2^j \boldsymbol{n} \in D_0} |W f(j, \boldsymbol{n}) - W f_i(j, \boldsymbol{n})|^2) - \sum_{j=1}^{L} \sum_{2^j \boldsymbol{n} \in D_1} |W f(j, \boldsymbol{n})|^2 \right\},$$

where D_0 is the set of pixels whose activity patterns are similar in all the source images, while D_1 is the set of pixels whose activity patterns are different. $\mathcal{F}$ is the set of all the images f whose wavelet transform satisfies

$$\min(Wf_i(j, \boldsymbol{n})) \leq Wf(j, \boldsymbol{n}) \leq \max(Wf_i(j, \boldsymbol{n})),$$

for any $0 < j \leq L$ and $\boldsymbol{n} \in \mathbb{Z}^2$. This constraint makes sure that the solution stays in the closure of $\mathcal{F}$, i.e., no image outside the scenario we are contemplating.

- **Shape recognition**

The geometrical description of an object can be decomposed into registration and shape information. For example, an object's location, rotation and size could be the registration information and the geometrical information that remains is the shape of the object. An object's shape is invariant under registration transformations and two objects have the same shape if they can be registered to match exactly.

The pioneers of this topic of general shape and registration analysis are Kendall [30] and Bookstein [31]. Some reference and reviews include Goodall [32], Kent [33], Dryden and Mardia [34].

The difficulty of target recognition in ISAR imagery is to identify a target shape in the presence of interference regardless of its position, scale and orientation. In an effort to overcome this difficulty, we describe matching of two configurations in Euclidean and affine shape spaces using regression techniques, and we further address robustness issues and matching by estimation.

1.2 Thesis Organization and Main Contributions

This section summarizes the organization of the remaining thesis, along with a brief description of the main contributions.

Chapter 2. Inverse Synthetic Aperture Radar Imaging System

Inverse Synthetic Aperture Radar System transmits electro-magnetic waves to a target and coherently integrates the returned signals to synthesize the effect of a larger aperture array. The spatial distribution of the reflectivity density of a target, referred to as the image of the target, is usually mapped onto a range-azimuth plain. In this chapter, we briefly introduce the range and azimuth processing, and necessary procedures to construct an ISAR image from reflected and measured signals.

Chapter 3. Inverse SAR Image Registration: An Information Theoretic Approach

In this chapter, a new generalized divergence measure, *Jensen-Rényi* divergence, is proposed. Some properties such as convexity and its upper bound are derived. Based on the Jensen-Rényi divergence, we propose a new approach to the problem of ISAR image registration. Our approach applies Jensen-Rényi divergence to measure the statistical dependence between consecutive ISAR image frames, which would be maximal if the images are geometrically aligned. Simulation results demonstrate that the proposed method is efficient and effective in tracking the trajectory of a target. The major results of this chapter have been published in [35], [36] and [37].

Chapter 4. Introduction to Multiscale Analysis

In this chapter, we briefly review the concept of multiscale analysis. We study the properties of an operator which approximates a signal at a given resolution. We show that the difference of a signal at different resolutions can be extracted by decomposing the signal on a wavelet orthonormal basis. The development of orthonormal wavelet bases has opened a new bridge between approximation theory and signal processing. In the last section of this chapter, we discuss a hard/soft threshold estimator and SureShrink in particular detail, and illustrate its adaptivity to unknown smoothness via wavelet shrinkage.

Chapter 5. Multiscale Signal Enhancement: Beyond the Normality and Independence Assumption

Most existing approaches to denoising or signal enhancement in a wavelet-based framework have generally relied on the assumption of normally distributed and independent

perturbations. In practice, this assumption is often violated and sometimes, even the prior information of a probability distribution of the noise process is unavailable. To relax this assumption, we propose a novel non-linear filtering technique in this chapter. The key idea is to project a noisy signal onto a wavelet domain and to suppress wavelet coefficients by a mask derived from its curvature extrema in a scale space representation. For a piecewise smooth signal, it can be shown that filtering by this curvature mask is equivalent to preserving the signal pointwise Hölder exponents at the singular points and lifting its smoothness at all the remaining points. The major results of this chapter have been published in [38] and [39].

Chapter 6. A Multiscale Approach for Pixel Level Image Fusion

Pixel level image fusion refers to the processing and synergistic combination of information gathered by various imaging sources to provide a better understanding of a scene. We formulate the image fusion as an optimization problem and propose an information theoretic approach in a multiscale framework to solve it. A biorthogonal wavelet transform of each source image is first calculated, and the new fusion algorithm applies *Jensen-Rényi* divergence to construct a composite of wavelet coefficients according to the measurement of the information patterns inherent in source images. Experimental results on fusion of multi-sensor navigation images, multi-modality medical images, multi-spectral remote sensing images, and multi-focus optical images are presented to illustrate the proposed fusion scheme. The major results of this chapter have been published in [40].

Chapter 7. Shape Recognition

The geometrical description of an object can be decomposed into registration and shape information. The goal of shape recognition is to identify a shape regardless of its registration information. In this chapter, we describe the matching of two configurations using a regression technique, making connections with general shape spaces and procrustean distances. In particular, we study the generalized matching by estimation in Euclidean and affine shape spaces. Simulation results show that matching by way of a mean shape is more robust than matching target shapes directly.

Chapter 8. Conclusions and Discussions

In this chapter, we briefly summarize contributions in this thesis and present the overall conclusions which can be drawn from the results of our research. We also present some suggestions for extending this work.

2 Inverse Synthetic Aperture Radar Imaging System
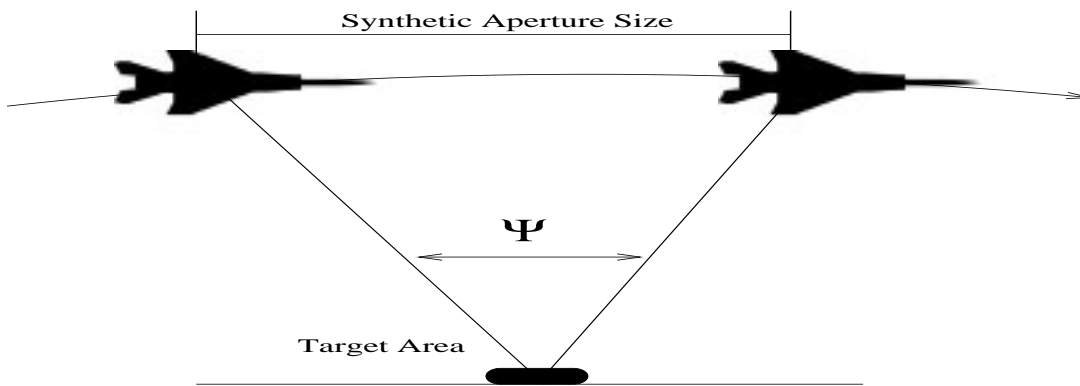
Figure 2.1: Spotlight SAR

INVERSE Synthetic Aperture Radar (ISAR) is a microwave imaging system capable of producing high resolution imagery from data collected by a relatively small antenna. ISAR can be explained in terms of spotlight SAR [6], as illustrated in Figure (2.1). Spotlight SAR is obtained as the radar antenna constantly tracks a particular target area of interest. The same data would be collected if the radar were stationary and the target area were rotating. The target rotation relative to the radar is used to generate the target image. This is precisely the idea of ISAR.

Figure (2.2) illustrates the data collection from an air target by a stationary radar as
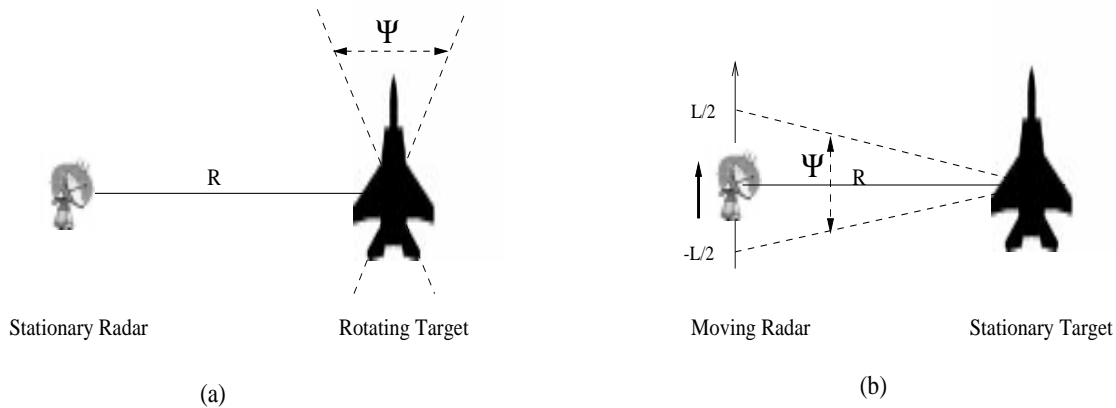
Figure 2.2: SAR/ISAR Equivalence: (a) ISAR; (b) SAR equivalence

the target rotates through an angle Ψ. The spotlight SAR equivalent geometry is the moving radar of Figure (2.2b), collecting the same data while flying a circular segment around an identical but non-rotating target. The SAR aperture length L in Figure (2.2b) corresponds to integration angle Ψ in figure (2.2a).

2.1   Range Processing

ISAR imagery represents reflectivity magnitude associated with the illuminated target. In the terminology of radar signal processing, the direction of radar Line of Sight (LOS) is referred to as *range* and the direction orthogonal to range is referred to as *cross-range* or *azimuth*.

Range is determined by measuring the time it takes for a transmitted signal to travel a round-trip distance between radar and target. The ability of the radar to determine the range of a particular scatterer in a vicinity of other scatterers depends on the range resolution.

Target reflectivity density is a function of frequency and viewing angle, and the assumption is that it does not vary significantly over the bandwidth of transmitted signal, which is considered narrow compared to the carrier frequency, or over the narrow

range of radar viewing angles.

Consider the simplified geometry shown in Figure (2.3). An antenna located at the origin illuminates a line of scatterers centered at $x = R$, having length W and reflectivity $q : \mathbb{R} \to \mathbb{C}$. Let the transmitted signal $\tilde{s}(t)$ be a pulse of duration T_p and bandwidth β given by

$$\tilde{s}(t) = \mathbf{Re}\{s(t)e^{j2\pi f_0 t}\}w(\frac{t}{T_p}) \tag{2.1}$$

where $s(t)$ is the baseband equivalent signal of $\tilde{s}(t)$, and

$$w(t) = \begin{cases} 1, & |t| \leq \frac{1}{2} \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

Then, ignoring the round-trip attenuation, the returned signal can be represented as the convolution of target reflectivity density with the transmitted signal

$$r(x) = q(x) \star s(2x/c) \tag{2.3}$$

where c is the speed of light, and $2x/c$ is the round trip delay.

An estimate of the reflectivity density can be obtained by passing $r(x)$ through a matched filter whose impulse response is $h_r(x) = s^*(-2x/c)$. Therefore, the estimate of $q(x)$ is

$$\begin{aligned} \hat{q}(x) &= r(x) \star s^*(-2x/c) \\ &= q(x) \star [s(2x/c) \star s^*(-2x/c)]. \end{aligned} \tag{2.4}$$

Let's define

$$\rho(t) = s(t) \star s^*(-t) \tag{2.5}$$

then we can rewrite (2.4) as

$$\hat{q}(x) = q(x) \star \rho(2x/c) \tag{2.6}$$

Since the transmitted pulse usually has a large time-bandwidth product, $\rho(t)$ can be approximated by a $\text{sinc}(\cdot)$ function. The estimate of target reflectivity density may thus be represented as

$$\hat{q}(x) = q(x) \star \text{sinc}(2\pi\beta x/c). \tag{2.7}$$

Figure 2.3: Simplified geometry for range processing.

The range resolution is hence given by [41]

$$\delta r = \frac{c}{2\beta} \quad .$$

(2.8)

2.2   Cross-Range Processing

To gain a basic understanding of the cross-range resolving mechanism in ISAR, let's consider Figure (2.4) with a line of scatterers having the reflectivity $q(y)$. As the radar is moving from $z = -L/2$ to $z = L/2$, the two-way phase advance at cross-range y is

$$\Psi_y(z) \approx \frac{4\pi}{\lambda} dr = \frac{4\pi}{\lambda} z \cdot \sin\gamma$$

(2.9)

For the case of $y \ll R$ and $L \ll R$, $\sin\gamma$ approaches y/R and z approaches $R\phi$, $-\Psi/2 < \phi < \Psi/2$, Equation (2.9) can be re-expressed in terms of angle ϕ

$$\Psi_y(\phi) \approx \frac{4\pi}{\lambda} R\phi \frac{y}{R} = \frac{4\pi}{\lambda} y\phi, \quad -\frac{\Psi}{2} < \phi < \frac{\Psi}{2}.$$

(2.10)

Then, the echo response from the line scatterers at ϕ is given by

$$g(\phi) = \int_{-\infty}^{+\infty} q(y) e^{j\frac{4\pi}{\lambda}\phi y} dy.$$

(2.11)

An estimate of the reflectivity density can be obtained by integrating the echo response
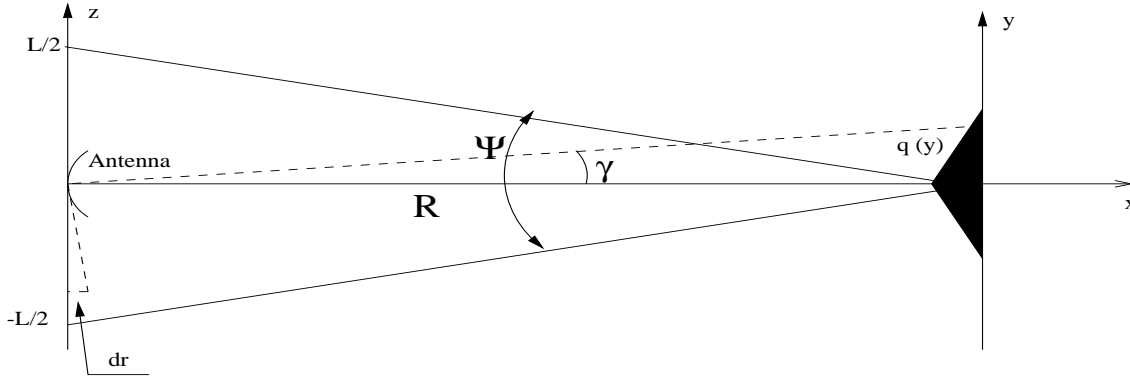
Figure 2.4: Simplified geometry for cross-range processing

over a small integration angle Ψ,

$$
\begin{aligned}
\hat{q}(y) &= \int_{-\frac{\Psi}{2}}^{+\frac{\Psi}{2}} g(\phi) e^{-j\frac{4\pi}{\lambda}\phi y} d\phi \\
&= \int_{-\infty}^{+\infty} g(\phi) w(\frac{\phi}{\Psi}) e^{-j\frac{4\pi}{\lambda}\phi y} d\phi \\
&= \Psi q(y) \star \operatorname{sinc}(\frac{2\pi}{\lambda}\Psi y),
\end{aligned}
\tag{2.12}
$$

where $w(\cdot)$ is the window function defined in equation (2.2). Therefore the cross-range resolution is given by [6]

$$
\delta r_c = \frac{\lambda}{2\Psi}.
\tag{2.13}
$$

2.3 Maximum Integration Angle

The maximum integration angle for ISAR imaging system before uncorrected rotation produces defocusing in cross-range can be defined in terms of allowable two-way phase deviation. We assume that the radar is in the far field of the target so that no significant range deviation exists over the target's cross-range extent. What remains is the range deviation produced when a scatterer first approaches and then recedes from the radar
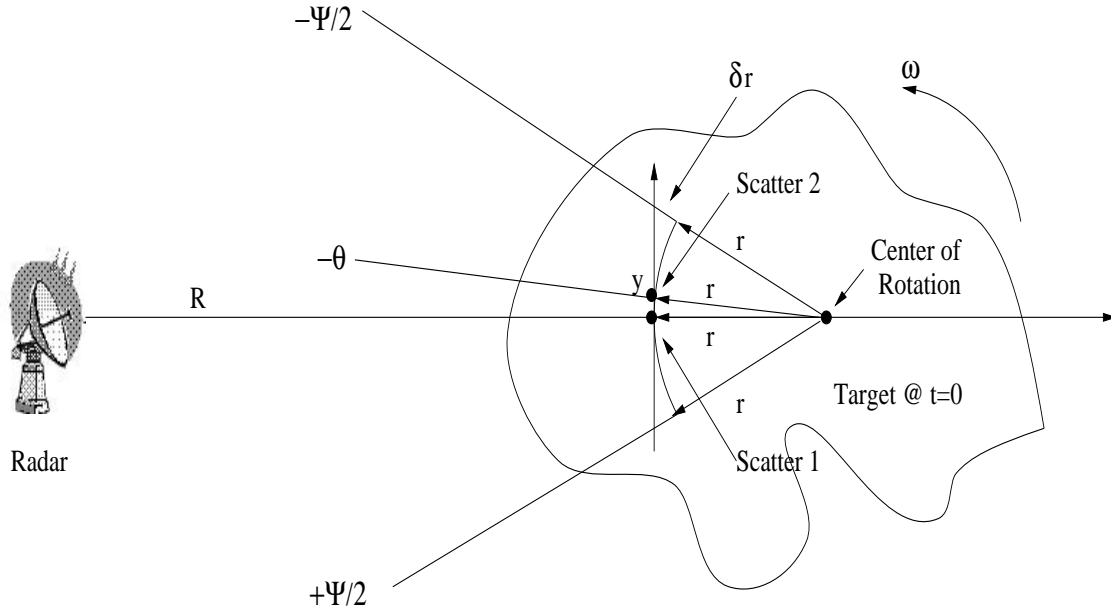
Figure 2.5: Unfocused ISAR geometry

during rotation. As in Figure (2.5), for a small Ψ,

$$r^2 + (\frac{r\Psi}{2})^2 = (r + \delta r)^2 \tag{2.14}$$

Solving for Ψ with $r >> \delta r$, we have

$$\Psi = (\frac{8\delta r}{r})^{1/2} \tag{2.15}$$

Assume a maximum allowable two-way phase deviation of $\pi/8$ as a criterion for focus, the corresponding allowable range deviation δr is $\lambda/32$. The maximum integration angle [6] then becomes

$$\Psi_{\max} = \frac{1}{2}\sqrt{\frac{\lambda}{r}}. \tag{2.16}$$

The cross-range resolution in unfocused ISAR is limited by the maximum integration angle $\Psi_{\max}$, and a *Focused Aperture* ISAR is desirable for a larger integration angle. Let's consider the echo signal from scatterer 1, located at $(r, \theta = 0)$ when $t = 0$ as shown in Figure (2.5), the phase advance term at time t is

$$\Psi_1(t) = -\frac{4\pi}{\lambda}\delta r = -\frac{4\pi}{\lambda}\frac{(\omega r t)^2}{2r} \tag{2.17}$$

where the target rotates at a constant angular rotation rate ω. The two way phase advance for scatterer 2, which is located at $(r, -\theta)$ when $t = 0$, is

$$\Psi_2(t) = -\frac{4\pi}{\lambda}\frac{(\omega t - \theta)^2 r}{2}. \tag{2.18}$$

A focused aperture for ISAR is achieved by subtracting $\Psi_1(t)$ from $\Psi_2(t)$, and we have

$$\Psi_y(t) = \Psi_2(t) - \Psi_1(t) = -\frac{4\pi}{\lambda}[-\omega t y + \frac{y^2}{2r}] \tag{2.19}$$

where $y = r\theta$. Define the target rotation angle $\phi = \omega t$, then we can rewrite Equation (2.19) in terms of ϕ,

$$\Psi_y(\phi) = -\frac{4\pi}{\lambda}[-y\phi + \frac{y^2}{2r}] \tag{2.20}$$

Equating (2.20) with (2.11), we obtain the range resolution [6]

$$\delta r_c = \frac{1}{2}\frac{\lambda}{\omega T}. \tag{2.21}$$

where T is the integration time. Equation (2.21) for focused ISAR with $\omega T = \Psi$ is the same as the cross-range resolution (2.13) for an unfocused, small integration angle ISAR.

The ISAR imaging is induced by target motion. During the imaging time, the scatterers must remain in their range cells. Reflectivity density function won't remain the same over a wide range of radar viewing angles. Therefore we can not use an arbitrary large integration angle in Equation (2.21). *Optimal Integration Angle* [35] need to be estimated to achieve the highest possible cross-range resolution and prevent defocusing in the image.

2.4   Range Doppler Imaging

The architecture of the ISAR image formation is illustrated in Figure (2.6), with a stepped-frequency (SF) waveform. Other wide-band radar waveforms, such as a linear FM chirp pulse can also be used but with different range compression techniques. Using
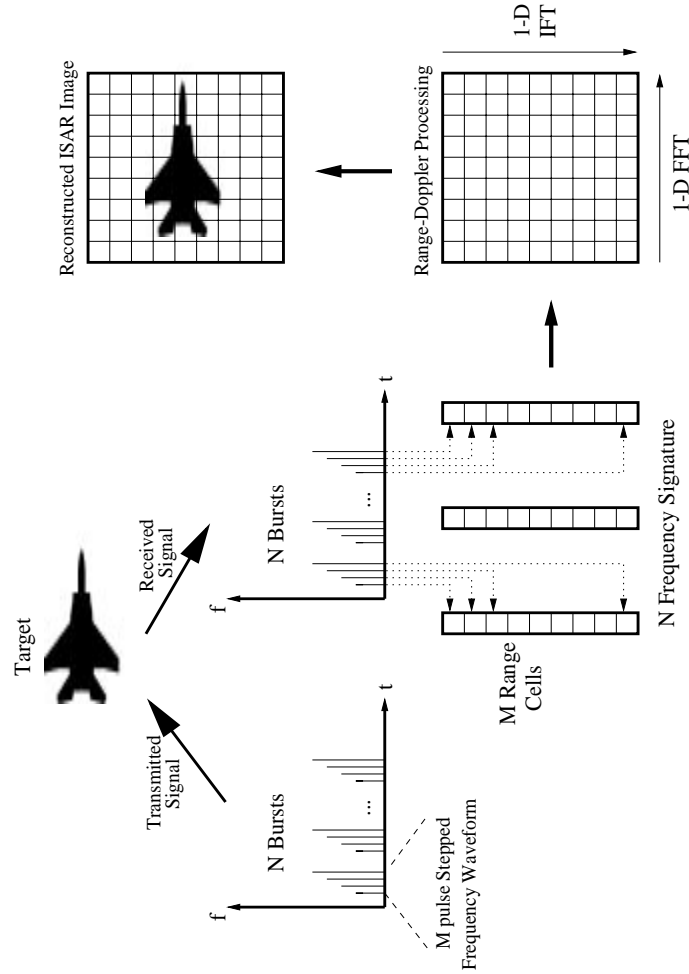
Figure 2.6: ISAR imaging system architecture

SF waveforms, a radar typically transmits a sequence of N bursts. Each burst consists of M narrow-band pulses. Within each burst, the center frequency of each successive pulse is increased by a constant frequency step Δf. The total bandwidth of the burst, i.e., $\beta = M\Delta f$, determines the radar range resolution. The total number of bursts for a given imaging time determines the Doppler or cross-range resolution.

To form a radar image, N bursts of returned signals first pass through a quadrature demodulator to obtain in-phase and quadrature signals at baseband. An $M \times N$ array of complex data, $\{G_{m,n}\}_{0 \leq m \leq M-1,\, 0 \leq n \leq N-1}$, is constructed to represent an unprocessed spatial frequency signature of the target. The radar processor uses the frequency signatures as the raw data to perform range compression and standard motion compensa-
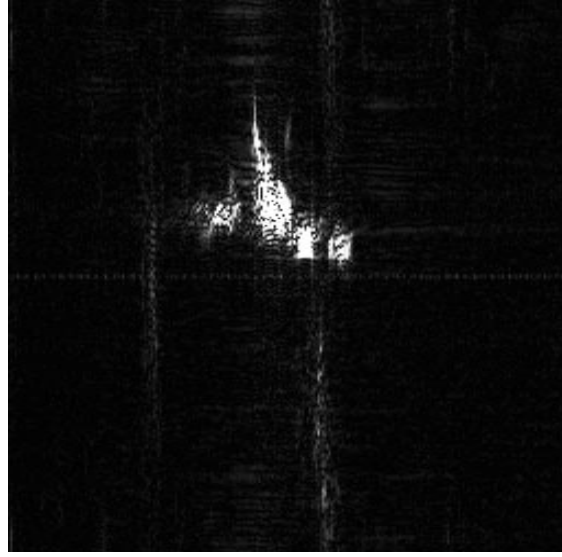
Figure 2.7: ISAR Image of A Small Boat

tion. Range compression functions as a matched filter to resolve range. For SF signals, the range compression performs an M-point IFT to each of the N received frequency signatures

$$r = \mathcal{F}_m^{-1}(G) \tag{2.22}$$

where $\mathcal{F}_m^{-1}$ denotes the IFT operation with respect to variable m. All together, N range profiles, each containing M range cells, are thus obtained.

Along each range cell, N range profiles constitute a time history series of the target. The Fourier imaging approach applies a FFT to the time history series and generates a N-point Doppler spectrum, or Doppler profile. By combining the N Doppler spectra at M range cells, a $M \times N$ image is formed

$$I = \mathcal{F}_n(r) \tag{2.23}$$

where $\mathcal{F}_n$ denotes a FFT operation with respect to variable n. The radar image I is hence the target's reflectivity function mapped onto the range-Doppler plane. Figure (2.7) shows such an ISAR image.

CHAPTER 3
Inverse SAR Image Registration: An Information Theoretic Approach

ENTROPY based divergence measures have shown promising results in many areas of engineering and image processing. In this paper, a new generalized divergence measure, *Jensen-Rényi* divergence, is proposed. Some properties such as convexity and its upper bound are derived. Based on the Jensen-Rényi divergence, we propose a new approach to the problem of ISAR (Inverse Synthetic Aperture Radar) image registration. The goal of ISAR image registration is to estimate the target motion during the imaging time. Our approach applies Jensen-Rényi divergence to measure the statistical dependence between consecutive ISAR image frames, which would be maximal if the images are geometrically aligned. Simulation results demonstrate that the proposed method is efficient and effective.

3.1 Introduction

Image registration is an important problem in computer vision [42] [43], remote sensing data processing [44] [45] and medical image analysis [46] [47]. The goal of image registration is to find a spatial transformation such that a similarity metric achieves its maximum between two or more images taken at different times, from different sensors, or from different viewpoints.

One such example which is the primary interest in the sequel is Inverse Synthetic Aperture Radar (ISAR) imaging. ISAR is a microwave imaging system capable of producing high resolution imagery from data collected by a relatively small antenna. ISAR imaging is induced by target motion, which unfortunately also blurs the resulting image. After a conventional ISAR translational focusing process, image registration can be applied to estimate the target rotational motion parameter, upon which polar re-formating may be used to achieve a higher resolution image. Related work in this area includes the image registration in interferometric SAR processing by Gabriel [48], Li [49] and Lin [50] and Fornaro [51].

Over the last three decades, a wide variety of registration techniques have been developed for different applications. These techniques can be classified [52] into correlation methods, Fourier methods, landmark mapping, and elastic model-based matching. Given two images $f_1, f_2 : I \times I \in \mathbb{R}^2 \rightarrow \mathbb{R}$, correlation methods [53] calculate the normalized two-dimensional cross-correlation function $C(f_1, \mathcal{T}_{(l,\theta,\gamma)} f_2)$ between f_1 and f_2, where $\mathcal{T}$ is a Euclidean transformation with translational parameter $l = (l_x, l_y)$, rotational parameter θ and scaling parameter γ. The registration problem may be succinctly stated as

$$(l^*, \theta^*, \gamma^*) = \arg \max_{l,\theta,\gamma} C(f_1, \mathcal{T}_{(l,\theta,\gamma)} f_2). \tag{3.1}$$

The correlation methods are generally limited to registration problems in which the image is misaligned by only a small rigid transformation. In addition, the peak of the correlation may not be clearly discernible in the presence of noise. Fourier methods [54] are the frequency domain equivalent of the correlation methods. Fourier methods make use of the translation property of the Fourier transform and search for the optimal spectrum matching between two images. Since rotation is invariant under a Fourier transformation, rotating an image merely rotates the Fourier transform of that image [55]. If we denote F_1, F_2 as the two-dimensional Fourier transform of f_1, f_2 respectively, we obtain the phase of the cross-power spectrum rotated by θ as

$$P_\theta(\omega_x, \omega_y) = \frac{F_1(\omega_x, \omega_y)\, \mathcal{T}_\theta F_2(\omega_x, \omega_y)}{|F_1(\omega_x, \omega_y)\, \mathcal{T}_\theta F_2(\omega_x, \omega_y)|}. \tag{3.2}$$

To determine the rotational parameter θ, one proceeds to maximize the two-dimensional inverse Fourier transformation of $P_\theta(\omega_x, \omega_y)$, i.e., a cross-correlation which is as peaked or as impulsive as possible, and the location of that impulse is exactly the translational parameter l. In light of their equivalence to the correlation methods, Fourier methods are also limited to registration problems with a small rigid transformation. If there exists spatially local variation, then both the correlation methods and the Fourier methods would fail. For cases of unknown misalignment type, landmark mapping techniques [56] and elastic model-based matching [57] [58] may be used to tackle the registration problem. Landmark mapping techniques extract feature points from a reference image and a target image respectively, then apply piecewise interpolation to compute a transformation to map the feature point sets from the reference image to the target image. Landmark based methods are usually labor intensive and their accuracy depends on the degree of reliability of the feature points. Instead of finding the mapping between the feature point sets, elastic model-based matching methods model the distortion in the image as the deformation of an elastic material. The resulting registration transformation is the deformation with a minimal bending and stretching energy. Practical elastic model-based methods [59] are based on computation expensive iterative algorithms, and the choice of feature points plays a crucial role in its performance.

In the work of Woods [1] and Viola [2], mutual information, a basic concept from information theory, is introduced as a measure of evaluating the similarity between images. When the two images are properly matched, corresponding areas overlap, and the resulting joint histogram contains high values for the pixel combinations of the corresponding regions. When the images are mis-registered, non-matched areas also overlap and will contribute to additional pixel combinations in the joint histogram. In case of misregistration, the joint histogram has less significant peaks and is more dispersed than that of the correct alignment of images. The registration criterion is hence to find a transformation such that the mutual information of the corresponding pixel pair intensity values in the matching images is maximized. This approach is accepted widely [3] as one of the most accurate and robust registration measures. Following the

same argument, Hero [4] *et al.* extends the concept to apply Rényi entropy to measure the joint histogram as a dissimilarity metric between images.

Inspired by this previous work, and looking to address their limitation in often difficult imagery, we introduce in this chapter a novel generalized information theoretic measure, namely the *Jensen-Rényi* divergence and defined in terms of Rényi entropy [5]. Jensen-Rényi divergence is defined as a similarity measurement among any finite number of weighted probability distributions. Shannon mutual information is a special case of the Jensen-Rényi divergence. This generalization provides us an ability to control the measurement sensitivity of the joint histogram, to ultimately result in a better registration accuracy.

In the next section, we give a brief statement of the problem. In section 3.3, we introduce the Jensen-Rényi divergence and its properties. In Section 3.4, we derive the performance upper bounds of the Jensen-Rényi divergence in terms of the Bayes error and also of the asymptotic error of the Nearest Neighbor Classifier. Section 3.5 describes the concepts of image registration with the Jensen-Rényi divergence. Numerical experiments for ISAR image registration is demonstrated in Section 3.6. Finally, we provide concluding remarks in Section 3.7.

3.2 Problem Statement

ISAR imagery is induced by target motion, and the target motion in turn causes time-varying spectra of the received signals. Motion compensation has to be applied to obtain a high resolution image. The objective of ISAR image registration is to estimate the target motion during the imaging time. Let $\mathcal{T}_{(l,\theta,\gamma)}$ be a Euclidean transformation with translational parameter $l = (l_x, l_y)$, rotational parameter θ and scaling parameter γ. Given two ISAR image frames f_1 and f_2, the estimates of target motion parameters $(l^*, \theta^*, \gamma^*)$ are given by

$$(l^*, \theta^*, \gamma^*) = \arg \max_{(l,\theta,\gamma)} D_{JR}\boldsymbol{\omega}_\alpha \left(f_1, \mathcal{T}_{(l,\theta,\gamma)} f_2 \right) \tag{3.3}$$

where $D_{JR_\alpha}\boldsymbol{\omega}(\cdot)$ is an induced similarity measure based on Jensen-Rényi divergence of order α and weight $\boldsymbol{\omega}$, which is maximal when f_1 matches $\mathcal{T}_{(l,\theta,\gamma)} f_2$. As the radar tracks the target, the reflected signal is continuously recorded during the imaging time. By registering a sequence of consecutive image frames, $\{f_i\}_{i=0}^N$, the target motion during the imaging time can be estimated by interpolating $\{(l_i, \theta_i, \gamma_i)\}_{i=1}^N$. Based on the estimated trajectory of the target, translational motion compensation (TMC), and rotational motion compensation (RMC) [6] can be used to generate a focused image of the target.

3.3   The Jensen-Rényi Divergence

Let $k \in \mathbb{N}$ and $\mathcal{X} = \{x_1, x_2, \ldots, x_k\}$ be a finite set with a probability distribution $\boldsymbol{p} = (p_1, p_2, \ldots, p_k)$, i.e. $\sum_{j=1}^k p_j = 1$ and $p_j = P(x_j) \geq 0$, where $P(\cdot)$ denotes the probability. Rényi entropy is a generalization of Shannon entropy [60], and is defined as

$$R_\alpha(\boldsymbol{p}) = \frac{1}{1-\alpha} \log \sum_{j=1}^k p_j^\alpha, \quad \alpha > 0 \text{ and } \alpha \neq 1. \tag{3.4}$$

For $\alpha > 1$, the Rényi entropy is neither concave nor convex.

For $\alpha \in (0, 1)$, it is easy to see that Rényi entropy is concave, and tends to Shannon entropy $H(\boldsymbol{p})$ as $\alpha \to 1$. It can easily be verified that R_α is a non-increasing function of α, and hence

$$R_\alpha(\boldsymbol{p}) \geq H(\boldsymbol{p}), \quad \forall \alpha \in (0, 1). \tag{3.5}$$

In the sequel, we will restrict $\alpha \in (0, 1)$, unless otherwise specified, and will use a base 2 logarithm, i.e., the measurement unit is *bits*.

As shown in Figure 3.1, uncertainty is at a minimum when Shannon entropy is used, and it increases as α decreases. Rényi entropy attains a maximum uncertainty when α is equal to zero.

Definition 3.1 *Let $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n$ be n probability distributions on $\mathcal{X}$ and $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_n)$ be a weight vector such that $\sum_{i=1}^n \omega_i = 1$ and $\omega_i \geq 0$. The Jensen-Rényi divergence is defined*
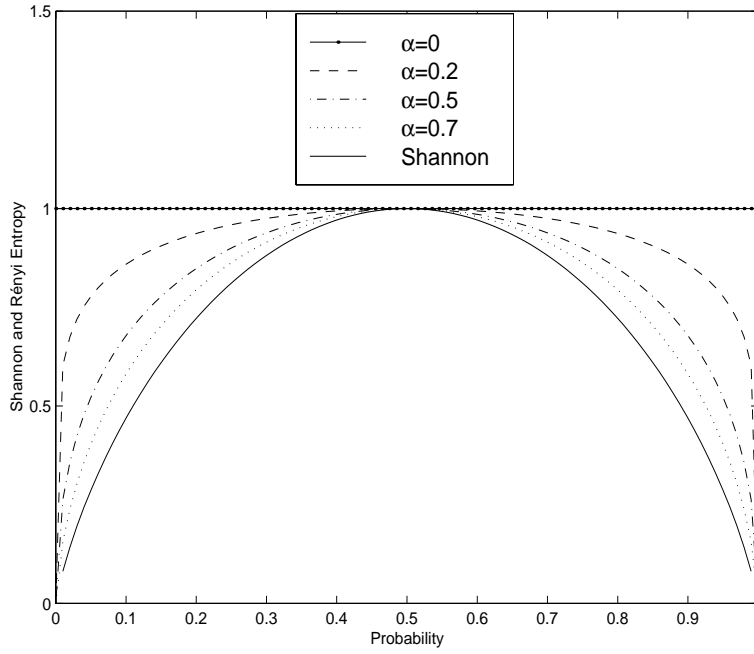
Figure 3.1: Shannon and Rényi entropy of Bernoulli distribution $\boldsymbol{p} = (p, 1-p)$ for different values of α

as

$$JR_\alpha^{\boldsymbol{\omega}}(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n) = R_\alpha\left(\sum_{i=1}^n \omega_i \boldsymbol{p}_i\right) - \sum_{i=1}^n \omega_i R_\alpha(\boldsymbol{p}_i),$$

where $R_\alpha(\boldsymbol{p})$ is the Rényi entropy, $\alpha > 0$ and $\alpha \neq 1$.

Using the Jensen inequality, it is easy to check that the Jensen-Rényi divergence is non-negative for $\alpha \in (0,1)$. It is also symmetric and vanishes if and only if the probability distributions $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n$ are equal, for all $\alpha > 0$. Figure 3.2 illustrates the three dimensional representation of the Jensen-Rényi divergence for two Bernoulli probability distributions, with $\alpha = 0.7$.

When $\alpha \to 1$, the Jensen-Rényi divergence is exactly the generalized Jensen-Shannon divergence [61].

Unlike other entropy-based divergence measures such as the well-known Kullback divergence, the Jensen-Rényi divergence has the advantage of being symmetric and generalizable to any finite number of probability distributions, with a possibility of assigning weights to these distributions.
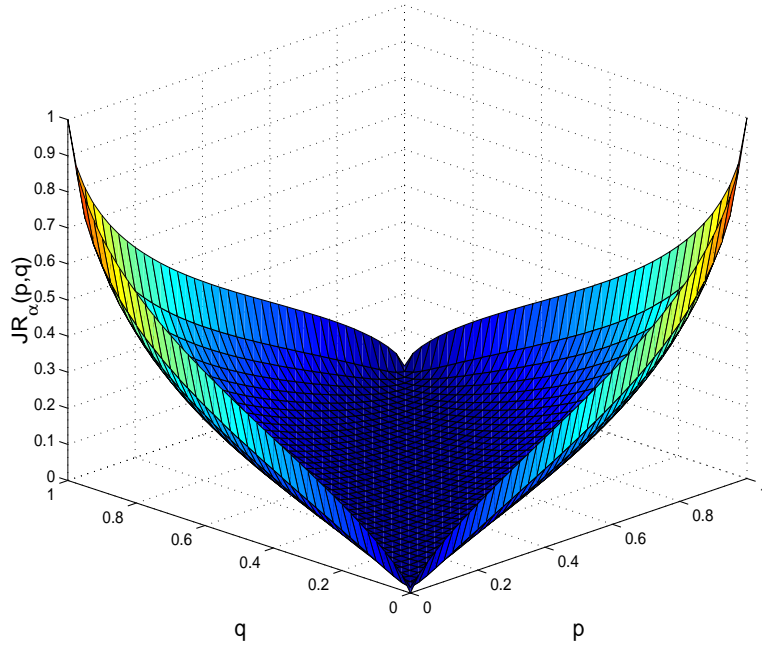
Figure 3.2: 3D representation of Jensen-Rényi divergence $JR_\alpha^{\boldsymbol{\omega}}(\boldsymbol{p}, \boldsymbol{q})$, $\boldsymbol{p} = (p, 1 - p)$, $\boldsymbol{q} = (q, 1 - q)$, $\alpha = 0.7$, $\boldsymbol{\omega} = (0.5, 0.5)$.

The following result establishes the convexity of the Jensen-Rényi divergence of a set of probability distributions.

Proposition 3.1 *For* $\alpha \in (0, 1)$, *the Jensen-Rényi divergence* $JR_\alpha^{\boldsymbol{\omega}}$ *is a convex function of* $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n$.

Proof: see Appendix.

The following result, in a sense, clarifies and justifies calling upon the Jensen-Rényi divergence as a measure of disparity among probability distributions.

Proposition 3.2 *The Jensen-Rényi divergence achieves its maximum value when* $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n$ *are degenerate distributions.*

Proof: The domain of $JR_\alpha^{\boldsymbol{\omega}}$ is a convex polytope in which the vertices are degenerate probability distributions. That is, the maximum value of the Jensen-Rényi divergence occurs at one of the degenerate distributions. ∎

Since the Jensen-Rényi divergence is a convex function of $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n$, it achieves its

maximum value when the Rényi entropy function of the ω-weighted average of degenerate probability distributions, achieves its maximum value as well.

Assigning weights ω_i to the degenerate distributions $\Delta_1, \Delta_2, \ldots, \Delta_n$, $\Delta_i = \{\delta_{ij}\}$, $j = 1, 2, \ldots k$, the following upper bound

$$JR_\alpha^{\boldsymbol{\omega}} \leq R_\alpha \left(\sum_{i=1}^n \omega_i \Delta_i \right), \tag{3.6}$$

which easily falls out of the Jensen-Rényi divergence, may be used as a starting point. Without loss of generality, consider the Jensen-Rényi divergence with equal weights $\omega_i = 1/n$ for all i, and denote it simply by JR_α, to write

$$\begin{aligned}
JR_\alpha &\leq R_\alpha \left(\sum_{i=1}^n (\Delta_i/n) \right) \\
&= \frac{1}{1-\alpha} \log \sum_{j=1}^k \left(\sum_{i=1}^n (\delta_{ij}/n) \right)^\alpha \\
&= R_\alpha(\boldsymbol{a}) + \frac{\alpha}{\alpha - 1} \log(n),
\end{aligned} \tag{3.7}$$

where

$$\boldsymbol{a} = (a_1, a_2, \ldots, a_k) \quad \text{such that} \quad a_j = \sum_{i=1}^n \delta_{ij}. \tag{3.8}$$

Since $\Delta_1, \Delta_2, \ldots, \Delta_n$ are degenerate distributions, then we have $\sum_{j=1}^k a_j = n$, $\forall k \geq n$. From (3.7), it is clear that JR_α achieves its maximum value when $R_\alpha(\boldsymbol{a})$ does as well. In order to maximize $R_\alpha(\boldsymbol{a})$, the concept of majorization [62] will be used. Let $(x_{[1]}, x_{[2]}, \ldots, x_{[k]})$ denote a non-increasing ordering of the components of a vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$.

Definition 3.2 *Let a and $b \in \mathbb{R}^k$, a is said to be majorized by b, written $a \prec b$, if*

$$\begin{cases} \sum_{j=1}^k a_{[j]} = \sum_{j=1}^k b_{[j]} \\ \sum_{j=1}^\ell a_{[j]} \leq \sum_{j=1}^\ell b_{[j]}, \quad \ell = 1, 2, \ldots, k-1. \end{cases}$$

Definition 3.3 *A real-valued function ϕ defined on a set $\Omega \subset \mathbb{R}^k$ is said to be Schur-concave on Ω if*

$$\boldsymbol{a} \prec \boldsymbol{b} \quad \Longrightarrow \quad \phi(\boldsymbol{a}) \geq \phi(\boldsymbol{b}), \quad \forall \boldsymbol{a}, \boldsymbol{b} \in \Omega.$$

Define the function $g(a_j) = a_j^\alpha$, $\alpha \in (0, 1)$, on an interval $J \subset \mathbb{R}$. It is clear that g is a concave function on J, thus $\phi(a) = \sum_{j=1}^k g(a_j)$ is Schur-concave [62] on J^k, that is

$$a \prec b \implies \sum_{j=1}^k g(a_j) \geq \sum_{j=1}^k g(b_j).$$

Since $\log(\cdot)$ is an increasing function, and $\alpha \in (0, 1)$, it follows that

$$a \prec b \implies R_\alpha(a) \geq R_\alpha(b).$$

Therefore, $R_\alpha(\cdot)$ is a Schur-concave function. The following result establishes the maximum value of the Jensen-Rényi divergence.

Proposition 3.3 *Let $p_1, p_2, \ldots, p_n$ be n probability distributions with*

$$p_i = (p_{i1}, p_{i2}, \ldots, p_{ik}), \quad \sum_{j=1}^k p_{ij} = 1, \quad p_{ij} \geq 0.$$

If $k \equiv r$ (mod n), $0 \leq r < n$, then

$$JR_\alpha \leq \frac{1}{1-\alpha} \log(k - r), \tag{3.9}$$

where $\alpha \in (0, 1)$.

Proof: It is clear that the vector

$$g = (\overbrace{n/(k - r), \ldots, n/(k - r)}^{k-r}, \overbrace{0, \ldots, 0}^{r})$$

is majorized by the vector a defined in (3.8). Hence, $R_\alpha(a) \leq R_\alpha(g)$. Invoking Equation (3.7) completes the proof. ∎

According to *Proposition 3.3*, and for the special case of $k \equiv 0$ (mod n) the following inequality holds

$$JR_\alpha(p_1, p_2, \ldots, p_n) \leq \log(k).$$

It is of special interest for the Jensen-Rényi divergence between the histograms p_a and p_b with weights $\{\beta, 1 - \beta\}$, $\beta \in [0, 1]$. Let

$$p = (1 - \beta/2)p_a + (\beta/2)p_b,$$

then the corresponding Jensen-Rényi divergence can then be expressed as a function of β

$$JR_\alpha(\beta) = R_\alpha(\boldsymbol{p}) - \frac{R_\alpha((1-\beta)\boldsymbol{p}_a + \beta\boldsymbol{p}_b) + R_\alpha(\boldsymbol{p}_a)}{2}.$$
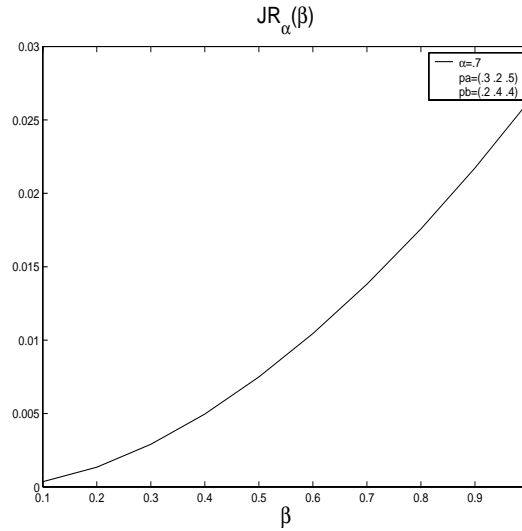
Figure 3.3: Jensen-Rényi divergence as a function of β

Proposition 3.4 *If $\boldsymbol{p}_a \neq \boldsymbol{p}_b$ then the Jensen-Rényi divergence $JR_\alpha(\beta)$ achieves its maximum value when $\beta = 1$.*

Proof: see Appendix.

Figure (3.3) illustrates the Jensen-Rényi divergence as a function of β when $\boldsymbol{p}_a = (0.3, 0.2, 0.5)$, $\boldsymbol{p}_b = (0.2, 0.4, 0.4)$ and $\alpha = 0.7$.

3.4   The Jensen–Rényi Divergence: Performance Bounds

In this section, performance bounds of the Jensen-Rényi divergence in terms of the Bayes error and also of the asymptotic error of the NN classifier are derived.

Let $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ...\boldsymbol{x}_k\}$ be a set of feature vectors and $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$ be a set of n classes. By a classifier we mean a function $f : \mathcal{X} \to \mathcal{C}$ that classifies a given feature

vector (pattern) $x \in \mathcal{X}$ to the class $c = f(x)$. Denote X, C be two random variables taking values in $\mathcal{X}$ and $\mathcal{C}$ respectively. It is well known that the classifier that minimizes the error probability $P(f(X) \neq C)$ is the result of the Bayes classifier with an error L_B written in discrete form as

$$L_B = \inf_{f:\mathcal{X} \to \mathcal{C}} P\{f(X) \neq C\} = 1 - \sum_{j=1}^{k} \max_{1 \leq i \leq n} \{\omega_i p_{ij}\},$$

where $\omega_i = P(C = c_i)$ are the class probabilities, and $p_{ij} = P(X = x_j | C = c_i)$ are the class-conditional probabilities. Denote by $\boldsymbol{\omega} = \{\omega_i\}_{1 \leq i \leq n}$, and $\boldsymbol{p}_i = (p_{ij})_{1 \leq j \leq k}$, $\forall i = 1, \ldots, n$.

Proposition 3.5 *The Jensen-Rényi divergence is upper bounded by*

$$JR_\alpha^{\boldsymbol{\omega}}(\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n) \leq R_\alpha(\boldsymbol{\omega}) - 2L_B, \tag{3.10}$$

where $R_\alpha(\boldsymbol{\omega}) = R_\alpha(C)$, and $\alpha \in (0, 1)$.

Proof: According to proposition 1, we have

$$R_\alpha(C) - JR_\alpha^{\boldsymbol{\omega}}(\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n) = R_\alpha(C|X).$$

It has been proven in [63] that $H(C|X) \geq 2L_B$, then (3.5) implies $R_\alpha(C|X) \geq 2L_B$. This completes the proof. ■

A method that provides an estimate for the Bayes error without requiring knowledge of the underlying class distributions is based on the NN classifier which assigns a test pattern to the class of its closest pattern according to some metric [64].
For n sufficiently large, the following result relating the Bayes error L_B and the asymptotic error L_{NN} of the NN classifier holds [64]

$$\frac{n-1}{n}\left(1 - \sqrt{1 - \frac{n}{n-1}L_{NN}}\right) \leq L_B \leq L_{NN}. \tag{3.11}$$

Using (3.10), the following inequality is deduced

$$JR_\alpha^{\boldsymbol{\omega}} \leq R_\alpha(\boldsymbol{\omega}) - \frac{2(n-1)}{n}\left(1 - \sqrt{1 - \frac{n}{n-1}L_{NN}}\right).$$

3.5   Image Registration with the Jensen-Rényi Divergence
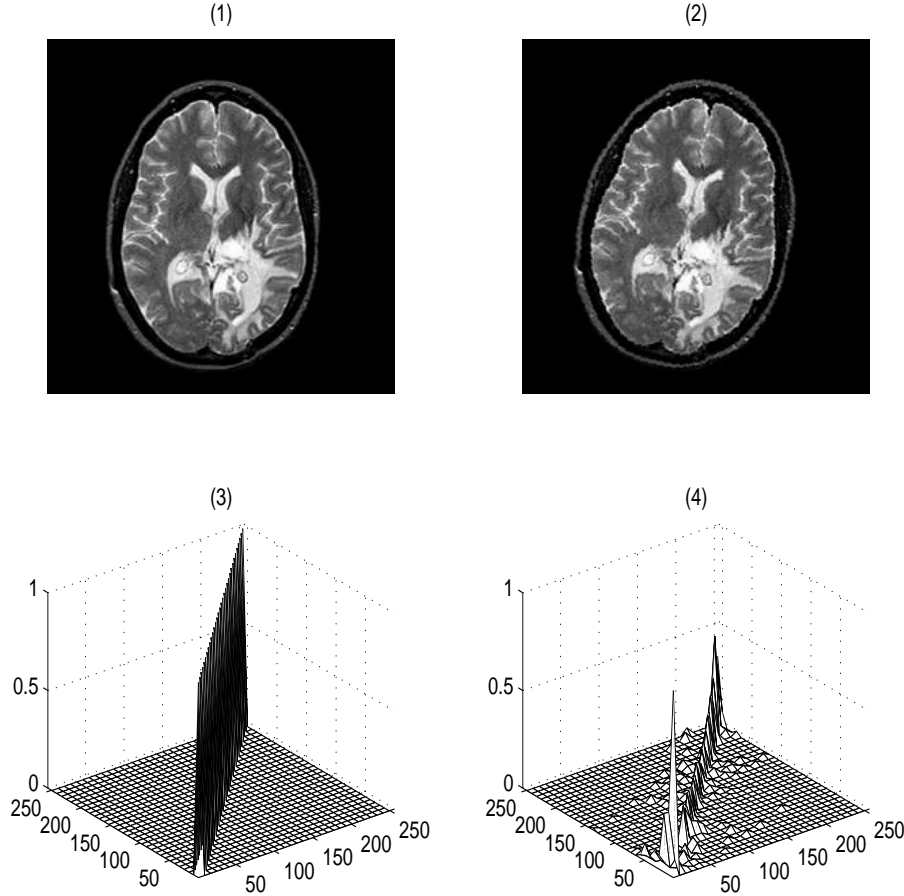
Figure 3.4: Conditional probability distributions

Let f_1, f_2 be two digital images defined on a bounded domain $\Omega \subseteq \mathbb{N}^2$, the goal of image registration is to determine the spatial transformation parameters $(l^*, \theta^*, \gamma^*)$ such that

$$
\begin{aligned}
(l^*, \theta^*, \gamma^*) &= \arg\max_{(l,\theta,\gamma)} D_{JR_\alpha^{\boldsymbol{\omega}}}(f_1, \mathcal{T}_{(l,\theta,\gamma)} f_2) \\
&= \arg\max_{(l,\theta,\gamma)} JR_\alpha^{\boldsymbol{\omega}}(\boldsymbol{p}_1(f_1, \mathcal{T}_{(l,\theta,\gamma)} f_2), \boldsymbol{p}_2(f_1, \mathcal{T}_{(l,\theta,\gamma)} f_2), \ldots, \boldsymbol{p}_n(f_1, \mathcal{T}_{(l,\theta,\gamma)} f_2))
\end{aligned}
\tag{3.12}
$$

where $JR_\alpha^{\boldsymbol{\omega}}(\cdot)$ is the Jensen-Rényi divergence of order α and weight $\boldsymbol{\omega}$.

Denote $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ and $\mathcal{Y} = \{y_1, y_2, \ldots, y_n\}$ the sets of pixel intensity values of f_1 and $\mathcal{T}_{(l,\theta,\gamma)} f_2$ respectively, and let X, Y be two random variables taking values in $\mathcal{X}$

and $\mathcal{Y}$. $\boldsymbol{p}_i(f_1, \mathcal{T}_{(l,\theta,\gamma)} f_2) = (p_{ij})_{1 \le j \le n}$ is defined as

$$p_{ij} = P(Y = y_j | X = x_i), \; j = 1, 2, \ldots, n$$

which is the conditional probability of $\mathcal{T}_{(l,\theta,\gamma)} f_2$ given f_1 for the corresponding pixel pairs. Here the Jensen-Rényi divergence acts as a similarity measure between images. If the two images are exactly matched, then $\boldsymbol{p}_i = (\delta_{ij})_{1 \le j \le n}, \; i = 1, 2, \ldots, n$. Since $\boldsymbol{p}'_i s$ are degenerate distributions, by Proposition 3.2, the Jensen-Rényi divergence is maximized for a fixed α and $\boldsymbol{\omega}$. Figure (3.4.1-3.4.2) show two brain MRT images in which the misalignment is a Euclidean rotation. The conditional probability distributions $\{\boldsymbol{p}_i\}$ are crisp, as in Figure (3.4.3), when the two images are aligned, and dispersed, as in Figure (3.4.4), when they are not matched.

It is worth noting that the maximization of the Jensen-Rényi divergence holds for any α and $\boldsymbol{\omega}$ such that $0 \le \alpha \le 1$ and $\omega_i \ge 0, \; \sum_i \omega_i = 1$. If we take $\alpha = 1$ and $\omega_i = P(X = x_i)$ then, by Proposition 3.1, the Jensen-Rényi divergence is exactly the Shannon mutual information. Indeed, the Jensen-Rényi divergence induced similarity measure provides a more general framework for the image registration problem.
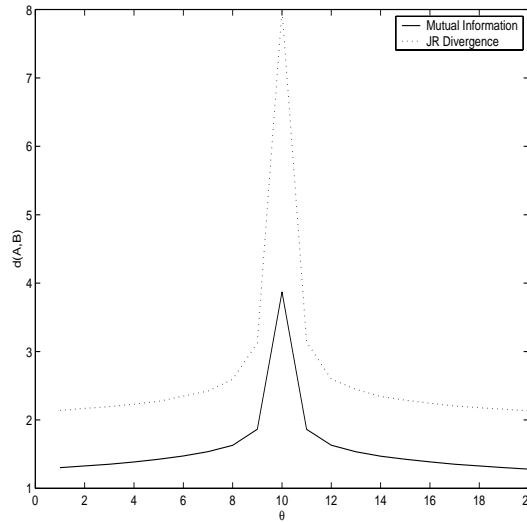
Figure 3.5: Mutual information vs. Jensen-Rényi divergence of uniform weights

If the two images f_1 and $\mathcal{T}_{(l,\theta,\gamma)} f_2$ are matched, the Jensen-Rényi divergence is maximized for any valid weight. Assigning $\omega_i = P(X = x_i)$ is not always a good choice.

Figure (3.5) shows the registration results of the two brain images in Figure (3.4) using the mutual information and the Jensen-Rényi divergence of $\alpha = 1$ and uniform weights. The peak at the matching point generated by the Jensen-Rényi divergence is clearly much higher than the peak by the mutual information. $\omega_i = P(X = x_i)$ gives the background pixel the largest weight. In the presence of noise, the matching in background is corrupted. Mutual information may fail to identify the registration point. This phenomenon is demonstrated in Figure (3.6). The following proposition establishes the optimality of the uniform weights for image registration in the context of the Jensen-Rényi divergence.

Proposition 3.6 *Let β be a uniform weight defined as $\beta_i = 1/n$, $i = 1, 2, \ldots, n$ and $\boldsymbol{\omega}$ be any vector such that $\omega_i \geq 0$, $\sum_{i=1}^{n} \omega_i = 1$. If the misalignment between f_1 and f_2 can be modeled by a spatial transformation $\mathcal{T}^*$, then the following inequality holds*

$$JR_\alpha^{\boldsymbol{\beta}}(\boldsymbol{p}_1(f_1, \mathcal{T}^* f_2), \ldots, \boldsymbol{p}_n(f_1, \mathcal{T}^* f_2)) \geq JR_\alpha^{\boldsymbol{\omega}}(\boldsymbol{p}_1(f_1, \mathcal{T}^* f_2), \ldots, \boldsymbol{p}_n(f_1, \mathcal{T}^* f_2)), \;\; \forall \alpha \in [0, 1].$$
(3.13)

Proof: $\boldsymbol{p}_i = \Delta_i$, $i = 1, 2, \ldots, n$ when f_1 and f_2 are aligned by the spatial transformation $\mathcal{T}^*$, then $JR_\alpha^{\boldsymbol{\omega}}(\cdot)$ becomes

$$JR_\alpha^{\boldsymbol{\omega}}(\boldsymbol{p}_1(f_1, \mathcal{T}^* f_2), \ldots, \boldsymbol{p}_n(f_1, \mathcal{T}^* f_2)) = R_\alpha(\sum_{i=1}^{n} \omega_i \Delta_i) = R_\alpha(\boldsymbol{\omega}).$$

Since $\beta \prec \omega$ [62] and $R_\alpha(\cdot)$ is Schur-concave, we obtain $R_\alpha(\boldsymbol{\beta}) \geq R_\alpha(\boldsymbol{\omega})$. This completes the proof of the proposition.   ∎

After assigning uniform weights to the various distributions in the Jensen-Rényi divergence, a free parameter α, which is directly related to the measurement sensitivity, remains to be selected. In the image registration problem, one desires a sharp and distinguishable peak at the matching point. The sharpness of the Jensen-Rényi divergence can be characterized by the maximal value as well as the width of the peak. The sharpest peak is clearly a Dirac function. The following proposition establishes that the maximal value of the Jensen-Rényi divergence is independent of α if the two images are aligned, and $\alpha = 0$ yields the sharpest peak.

(1) Image A   (2) Image B   (3) $d(A, B)$
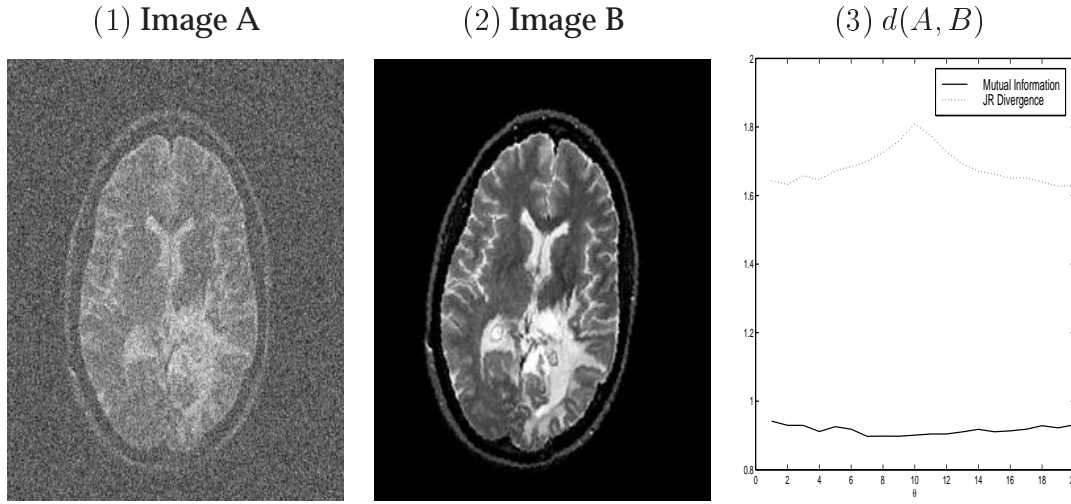
Figure 3.6: Registration result in the presence of the noise with $\alpha = 1$.

Proposition 3.7 *Let β be a uniform weight vector. If the misalignment between f_1 and f_2 can be modeled by a spatial transformation $\mathcal{T}^*$, then*

$$JR_\alpha^\beta \left(\boldsymbol{p}_1(f_1, \mathcal{T}^* f_2), \ldots, \boldsymbol{p}_n(f_1, \mathcal{T}^* f_2) \right) = \log n, \quad \forall \alpha \in [0, 1]. \tag{3.14}$$

In case of $\alpha = 0$,

$$JR_\alpha^\beta \left(\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n \right) = 0$$

for any probability distribution $\boldsymbol{p}_i$ such that $p_{ij} > 0, \ i, j = 1, 2, \ldots, n$ and

$$JR_\alpha^\beta \left(\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n \right) = \log n$$

if and only if $\boldsymbol{p}_i = \Delta_i, \ i = 1, 2, \ldots, n.$

Proof: see Appendix

If there exists local variation between f_1 and f_2, or, if the registration of the two images is in the presence of noise, then an exact alignment $\mathcal{T}^*$ may not be found. The conditional probability distribution $\boldsymbol{p}_i(f_1, \mathcal{T}^* f_2)$ is no longer a degenerate distribution in this case. The following proposition establishes that taking $\alpha = 1$ would provide a higher peak than any other choice of α for the non-ideal alignment.

Proposition 3.8 *Let $\boldsymbol{p}_i = \Delta_i + \delta\boldsymbol{p}_i$, $i = 1, 2, \ldots, n$, where $\delta\boldsymbol{p}_i = (\delta p_{ij})_{1 \leq j \leq n}$ is a real distortion vector such that $p_{ij} \geq 0$, $\sum_{j=1}^{n} \delta p_{ij} = 0$ and $\sum_{i=1}^{n} \delta p_{ij} = 0$. Let $\boldsymbol{\omega}$ be a weight vector and denote $JR^{\boldsymbol{\omega}}(\cdot)$ as the Jensen-Rényi divergence with $\alpha = 1$, then we have*

$$JR^{\boldsymbol{\omega}}(\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n) \geq JR^{\boldsymbol{\omega}}_{\alpha}(\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n), \quad \forall \alpha \in (0, 1). \tag{3.15}$$

Proof: Observe that for any probability distribution $\boldsymbol{p}$, $R_\alpha(\boldsymbol{p}) \geq H(\boldsymbol{p})$, $\forall \alpha \in (0, 1)$, then,

$$\sum_{i=1}^{n} \omega_i H(\boldsymbol{p}_i) \leq \sum_{i=1}^{n} \omega_i R_\alpha(\boldsymbol{p}_i), \quad \forall \alpha \in (0, 1) \tag{3.16}$$

Since $\sum_{j=1}^{n} \delta p_{ij} = 0$, $\sum_{i=1}^{n} \delta p_{ij} = 0$, and the Rényi entropy of $\alpha = 1$ is exactly the Shannon entropy, the inequality (3.15) is equivalent to the inequality (3.16). This completes the proof of *Proposition 3.8*. ∎

3.5.1 Discussion
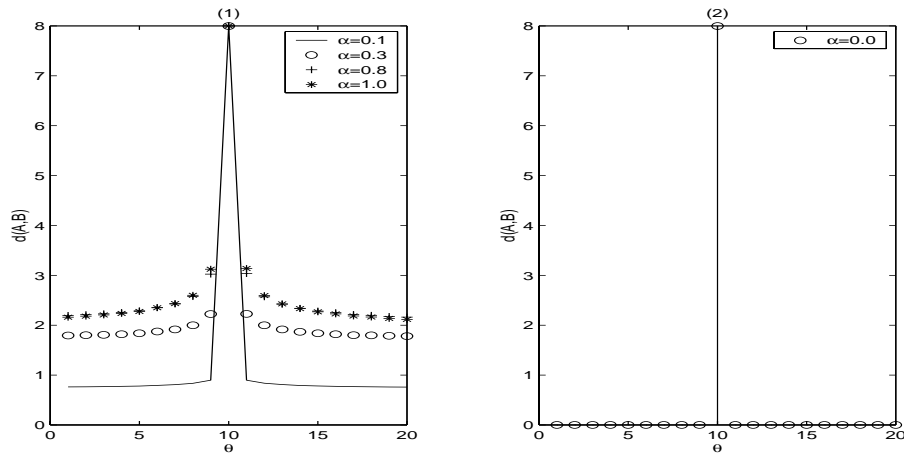
Figure 3.7: Effect of the order α in image registration

In real world applications, there is a trade off between optimality and practicality in choosing α. If one can model the misalignment between f_1 and f_2 completely and accurately, $\alpha = 0$ would correspond to the best choice since it generates the sharpest peak at the matching point. It is, however, also the least robust selection, as it tends to make

all the $p'_i s$ the same as the uniform distribution, if p_i is not degenerate distribution and $p_{ij} > 0$, then the Jensen-Rényi divergence would be zero for the whole transformation parameter space as in case where the adapted transformation group can not "accurately"[1] model the relationship between f_1 and f_2 accurate enough. On the other hand, $\alpha = 1$ is the most robust choice, in spite of also resulting in the least sharp peak. The choice of α therefore depends largely on the accuracy of the invoked model and on the specific application as well as the available computational resource. As an example, Figure (3.7.1) demonstrates the registration results of the two brain images in Figure (3.4) with the choice of different α. In this case, $\alpha = 0$ is the best choice and would generate a Dirac function with a peak at the matching point, as illustrated in Figure (3.7.2).

3.6 Numerical Experiments: ISAR Image Registration
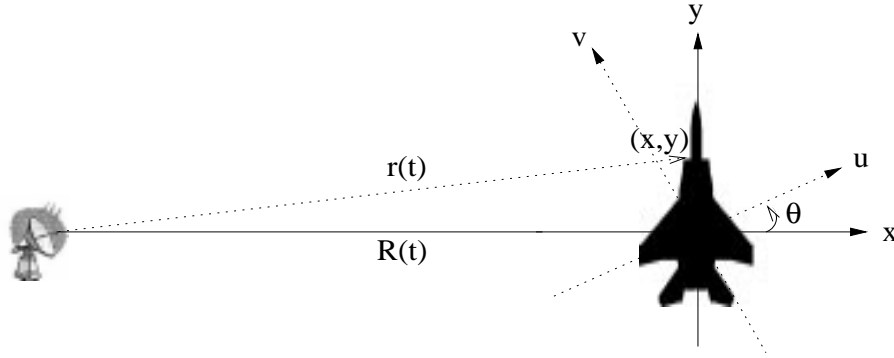
Figure 3.8: ISAR geometry of a moving target

Generating an ISAR image by using stepped frequency waveform [6] can be understood as a process of estimating the target's two-dimensional reflectivity density function $\rho(x, y)$ from data collected in the frequency space. Suppose a stepped frequency burst consists of M pulses in which the transmitted frequency linearly increases from ω_0 to $\omega_0 + (M - 1)\Delta\omega$, where ω_0 is the base frequency in *rads/s* and $\Delta\omega$ is the step frequency. Let the mth transmitted pulse $s_m(t)$ be a pulse of duration T_p and expressed in

---
[1]within some acceptable tolerance

complex form as

$$s_m(t) = e^{j\omega_m t} W\left(\frac{t - mT_p}{T_p}\right), \quad m = 0, 1, ...M - 1. \tag{3.17}$$

where $\omega_m = \omega_0 + m\Delta\omega$, and

$$W(t) = \begin{cases} 1, & 0 \leq t < 1 \\ 0, & \text{otherwise}. \end{cases} \tag{3.18}$$
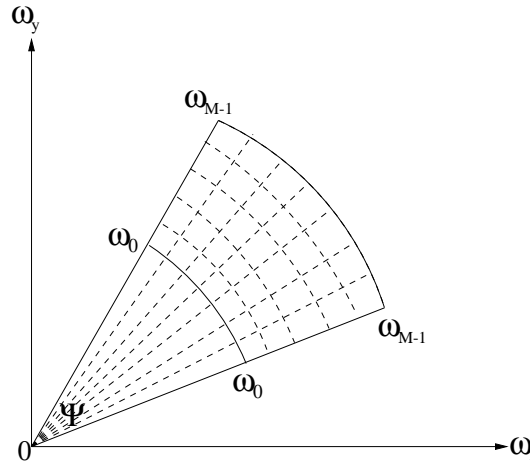
Figure 3.9: Polar formatted data in spatial frequency space

Let's define $\omega(t) = \omega_0 + (m - 1)\Delta\omega$, $mT_p \leq t < (m + 1)T_p$, $m = 0, 1, ...M - 1$. Under uniform illumination, the reflected signal from the target differential area $dx \times dy$ at the target coordinate (x, y) is

$$h(x, y, t) = A\rho(x, y)e^{j\omega(t)(t - 2r(t)/c)}dxdy, \quad 0 \leq t < (M - 1)T_p \tag{3.19}$$

where A is a constant attenuation factor, which we can set to 1 without a loss of generality. The distance between the radar antenna and the target reflection point located at (x, y) is denoted by $r(t)$. We obtain the expression of the received signal for $0 \leq t < (M - 1)T_p$ by integrating reflections from all the point scatterers in the target

$$\begin{aligned} \tilde{g}(t) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h(x, y, t)dxdy \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \rho(x, y)e^{j\omega(t)(t - 2r(t)/c)}dxdy. \end{aligned} \tag{3.20}$$

After quadrature demodulation, we obtain

$$g(t) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \rho(x,y) e^{-j2\omega(t)r(t)/c} dx dy., \ \ 0 \leq t < (M-1)T_p \tag{3.21}$$

It can be observed from Figure (3.8) that, for target dimension that are relatively smaller than the target range R, the distance $r(\cdot)$ from the radar antenna to target reflection point located at (x,y) is

$$r(t) \approx R(t) + x \cos \theta(t) - y \sin \theta(t). \tag{3.22}$$

Inserting Equation (3.22) into Equation (3.21), we deduce the baseband signal in terms of target coordinate (x,y) and rotation angle θ

$$g(t) = e^{-j2\omega(t)R(t)/c} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \rho(x,y) e^{-j(x\omega_x(t) - y\omega_y(t))} dx dy, \ \ 0 \leq t < (M-1)T_p \tag{3.23}$$

where

$$\omega_x(t) = \frac{2\omega(t)}{c} \cos \theta(t) \tag{3.24}$$

and

$$\omega_y(t) = \frac{2\omega(t)}{c} \sin \theta(t) \tag{3.25}$$

are spatial frequency quantities defined at frequency $\omega(t)$ and target rotation angle $\theta(t)$. The phase term $e^{-j2\omega(t)R(t)/c}$ is related to the target translational motion only, and can by compensated by traditional translational motion compensation methods.

By sampling $e^{j2\omega(t)R(t)/c} g(t)$ at $t_m = (m + \frac{1}{2})T_p, \ m = 0, 1, ...(M-1)$, we obtain the data collected in the frequency space $G(m)$ as

$$G(m) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \rho(x,y) e^{-j(x\omega_x(t_m) - y\omega_y(t_m))} dx dy, \ \ m = 0, 1, ...(M-1) \tag{3.26}$$

To form a radar image, N bursts of received signal are sampled and organized burst by burst into a $M \times N$ two-dimensional array, which is shown in Figure (3.9). This sample matrix is not uniformly spaced in the spatial frequency, instead, it is polar formatted data. The Discrete Fourier Transform processing of the polar formatted data would result in blurring at the edges of the target reflectivity image. Figure (3.10) is a synthetic
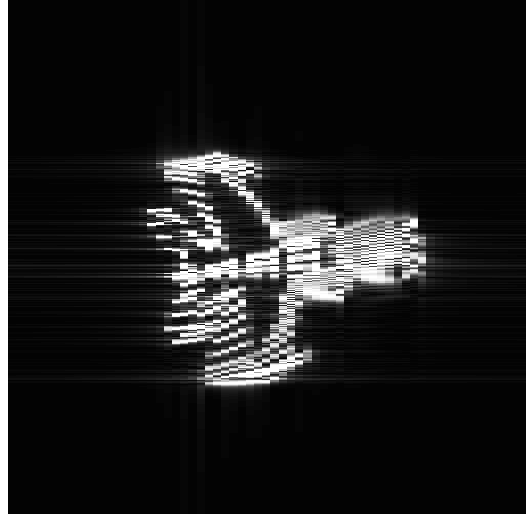
Figure 3.10: ISAR image of moving target reconstructed by the Discrete Fourier Transformation

ISAR image of an aircraft MIG-25 [65]. The radar is assumed to be operating at $9GHz$ and transmits a stepped-frequency waveform. Each burst consists of 64 narrow-band pulses stepped in frequency from pulse to pulse by a fixed frequency step of $8MHz$. The pulse repetition frequency is $15KHz$. Basic motion compensation processing has been applied to the data. A total of 512 bursts of received signal are taken to reconstruct the image of this aircraft, which corresponds to $2.18s$ integration time. As we can see, the resulting image is defocused due to the target rotation. In fact, the defocused image in Figure (3.10) is formed by overlapping a series of MIG-25s at different viewing angles. By replacing the Fourier transform with the time varying spectral analysis techniques [35] [65], we can take a sequence of snapshots of the target during the $2.18s$ of integration time. Figure (3.11.1-3.11.6) shows the trajectory of the MIG-25, with 6 image frames taken at $t = 0.1280s$, $0.4693s$, $0.8107s$, $1.1520s$, $1.4933s$, $1.8347s$ respectively.

Image registration can be applied to estimate the target motion from this sequence of images. For the synthetic ISAR images shown in Figure(3.11), we search for the rotation angles $\{\theta_i\}_{i=1}^N$ between a sequence of image frames $\{I_i\}_{i=0}^N$ observed in a time interval
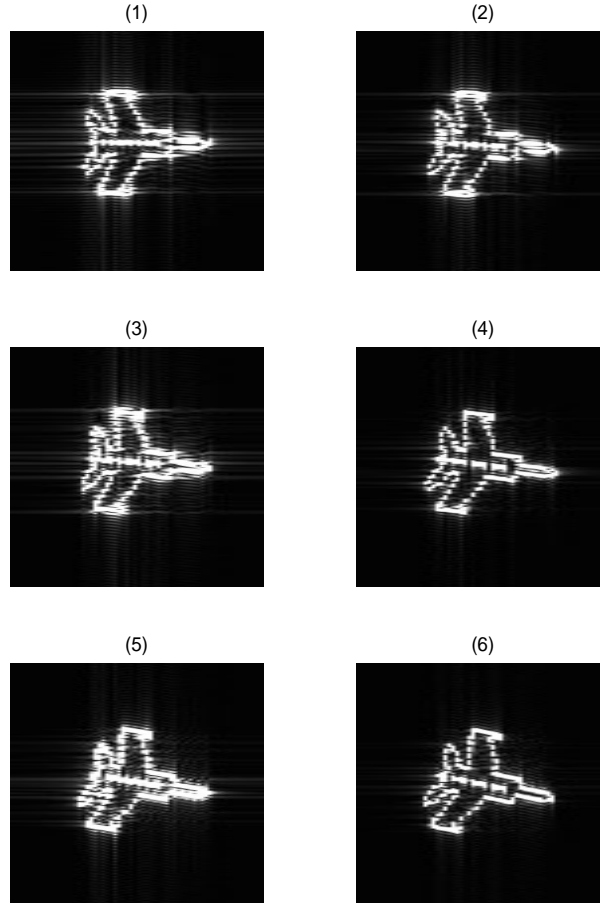
Figure 3.11: Trajectory of a sequence of MIG-25 image frames

$[0, T]$. By Equation (3.12), θ_i is given by

$$\theta_i^* = \arg \max_{\theta_i} JR_\alpha^{\boldsymbol{\omega}}(\boldsymbol{p}_1(I_{i-1}, \mathcal{T}_{\theta_i}I_i), \ldots, \boldsymbol{p}_n(I_{i-1}, \mathcal{T}_{\theta_i}I_i)).$$

Figure (3.12.1-3.12.5) shows the rotation angles $\{\theta_i\}_{i=1}^N$ obtained by registering the 6 consecutive MIG-25 image frames. As can already be seen in the figures, uniform weights produce the sharpest peak.

By interpolating $\{\theta_i\}_{i=1}^N$, we obtain an estimated trajectory of the MIG-25 rotational motion during the imaging time, as shown by the solid line in Figure (3.12.6). The dotted line in Figure (3.12.6) is the true trajectory. The standard deviation is $0.5580°$. An estimated trajectory of a target is particularly important since it may be subsequently used in polar re-formating [6] and re-sampling the received signal into rectangular format.
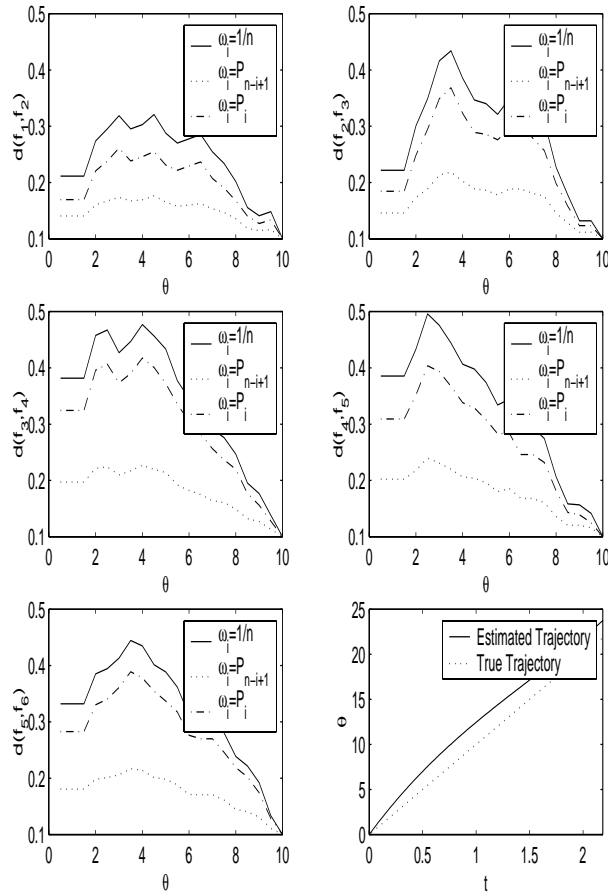
Figure 3.12: Image registration of a MIG-25 Trajectory

This results in a focused image of the MIG-25 based on all the received signals in the time interval $[0, \; 2.18s]$, as demonstrated in Figure (3.13).

3.7   Conclusions

A new generalized divergence measure, *Jensen-Rényi* divergence, is proposed in this paper. We prove the convexity of this divergence measure, derive its maximum value, and analyze its performance upper bounds in terms of the Bayes error of nearest neighbor classifier. Based on the Jensen-Rényi divergence, we propose a new approach to the problem of image registration. Compared to the mutual information based registration techniques, the Jensen-Rényi divergence adjusts its weight and exponential order to

control the measurement sensitivity of the joint histogram. This flexibility ultimately results in a better registration accuracy.
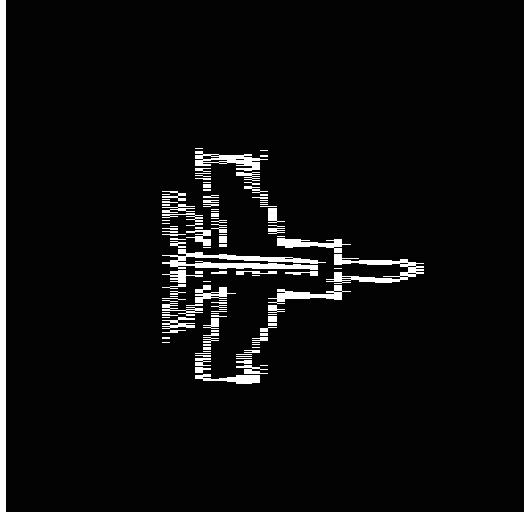
Figure 3.13: Reconstructed MIG-25 by polar reformatting

3.8 Appendix

Proof of Proposition 3.1

Denote $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ and $\mathcal{Y} = \{y_1, y_2, \ldots, y_n\}$. Let X, Y be two random variables taking values in $\mathcal{X}$ and $\mathcal{Y}$. Recall that the mutual information between X and Y is given by [66]

$$I(X;Y) = H(Y) - H(Y|X), \tag{3.27}$$

where $H(Y)$ is the Shannon entropy of Y and $H(Y|X)$ is the conditional Shannon entropy of Y given X.

Instead of using Shannon entropy in (3.27), the mutual information can be generalized using Rényi entropy. Therefore, the α-mutual information can be defined as

$$I_\alpha(X;Y) = R_\alpha(Y) - R_\alpha(Y|X), \quad \alpha \in (0, 1),$$

where R_α is the Rényi entropy of order $\alpha \in (0, 1)$.

Denote by $P(x_i) = \omega_i$, $P(Y = y_j | X = x_i) = p_{ij}$ and $P(Y = y_j) = q_j$, then it is easy to check that

$$R_\alpha(Y) - R_\alpha(Y|X) = JR_\alpha^{\boldsymbol{\omega}}(\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n), \tag{3.28}$$

where $\boldsymbol{p}_i = (p_{ij})_{1 \leq j \leq k}$, for all $i = 1, \ldots, n$.

For fixed ω_i, the mutual information is a convex function of p_{ij} [66], then it can be verified that the α-mutual information is also a convex function of p_{ij}, leading to the Jensen-Rényi divergence a convex function of $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n$. ■

Proof of Proposition 3.4

Denote by $\boldsymbol{p}_a = (p_{ak})_{k=0}^{L-1}$, $\boldsymbol{p}_b = (p_{bk})_{k=0}^{L-1}$ and $d_k = p_{bk} - p_{ak}$, where $L - 1$ is the maximum gray level of the image. The Jensen-Rényi divergence can then be written as

$$
\begin{aligned}
JR_\alpha(\beta) &= \frac{1}{1-\alpha} \log \sum_{k=0}^{L-1} \left(p_{ak} + \frac{\beta}{2} d_k \right)^\alpha \\
&\quad - \frac{1}{2(1-\alpha)} \log \sum_{k=0}^{L-1} (p_{ak} + \beta d_k)^\alpha \\
&\quad - \frac{1}{2(1-\alpha)} \log \sum_{k=0}^{L-1} p_{ak}^\alpha
\end{aligned}
$$

Simple calculations show that the first derivative $JR_\alpha'(\beta)$ vanishes for $\beta = 0$, and the second derivative $JR_\alpha''(\beta)$ is positive, that is $JR_\alpha'(\beta)$ is an increasing function of β. It follows that $JR_\alpha'(\beta) \geq 0$ for all $\beta \in [0, 1]$. Therefore, $JR_\alpha(\beta)$ is an increasing function of β. This concludes the proof. ■

Proof of Proposition 3.7

By Proposition 3.6,

$$
\begin{aligned}
& JR_\alpha^{\boldsymbol{\beta}} (\boldsymbol{p}_1(f_1, \mathcal{T}^* f_2), \ldots, \boldsymbol{p}_n(f_1, \mathcal{T}^* f_2)) \\
&= R_\alpha \left(\sum_{i=1}^{n} \beta_i \Delta_i \right) \\
&= \frac{1}{1-\alpha} \log \sum_{i=1}^{n} \left(\frac{1}{n} \right)^\alpha \\
&= \log n
\end{aligned}
$$

For $\alpha = 0$,

$$R_\alpha(\boldsymbol{p}_i) = \frac{1}{1-\alpha} \log \sum_{i=1}^{n} (p_{ij})^\alpha = \log n, \ \ \forall p_{ij} > 0, \sum_{j=1}^{n} p_{ij} = 1$$

we obtain

$$JR_\alpha^\beta(\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n) = R_\alpha(\sum_{i=1}^{n} \frac{1}{n}\boldsymbol{p}_i) - \sum_{i=1}^{n} \frac{1}{n}R_\alpha(\boldsymbol{p}_i) = 0.$$

If $\alpha = 0$ and $\boldsymbol{p}_i = \Delta_i$, $i = 1, 2, \ldots, n$, denote $0^0 = 0$, then by Proposition 3.6,

$$JR_\alpha^\beta(\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n) = \log n.$$

This concludes the proof for the proposition. ∎

4 Introduction to Multi-scale Analysis

IN this section, we briefly review the concept of multiscale analysis [67]. We study the properties of the operator which approximates a signal at a given resolution. We show that the difference of a signal at different resolutions can be extracted by decomposing the signal on a wavelet orthonormal basis. In $\mathrm{L}^2(\mathbb{R})$, a wavelet orthonormal basis is a family of functions $\{\sqrt{2^{-j}}\psi(2^{-j}t - n)\}_{(j,n)\in\mathbb{Z}^2}$, which is built by dilating and translating a unique function $\psi(\cdot)$. The development of orthonormal wavelet bases has opened a new bridge between approximation theory and signal processing. For the application of signal estimation in additive noise environment, linear operators have long predominated because of their simplicity, despite their limited performance. It is possible to keep the simplicity while improving the performance with non-linearities in a sparse representation. One such example is a wavelet thresholding estimator, which is used to suppress additive noises and restore signals degraded by low-pass filters.

4.1 Multiscale Approximation of $\mathrm{L}^2(\mathbb{R})$

Adapting the signal resolution allows one to process only the relevant details for a particular task. Burt and Adelson [68] introduced a multiresolution pyramid that can be used to process a low resolution first and then selectively increase the resolution when

necessary. The approximation of a function f at a resolution 2^{-j} is specified by samples on a discrete grid which provides local averages of f over neighborhoods of size proportional to 2^j.

Definition 4.1 *Let $\{\mathbf{V}_j\}, j \in \mathbb{Z}$ be a set of closed subspaces of $\mathbf{L}^2(\mathbb{R})$, we say that the sequence $(\mathbf{V}_j), j \in \mathbb{Z}$ is a multiresolution approximation of $\mathbf{L}^2(\mathbb{R})$, if the following conditions hold:*

$$\bullet \quad \forall (j, n) \in \mathbb{Z}^2, f(t) \in \mathbf{V}_j \iff f(t - 2^j n) \in \mathbf{V}_{2^j} \tag{4.1}$$

$$\bullet \quad \forall j \in \mathbb{Z}, \mathbf{V}_{j+1} \subset \mathbf{V}_j \tag{4.2}$$

$$\bullet \forall j \in \mathbb{Z}, f(t) \in \mathbf{V}_j \iff f(\frac{t}{2}) \in \mathbf{V}_{j+1} \tag{4.3}$$

$$\bullet \quad \lim_{j \to +\infty} \mathbf{V}_j = \bigcap_{j=-\infty}^{+\infty} \mathbf{V}_j = \{0\} \tag{4.4}$$

$$\bullet \quad \lim_{j \to -\infty} \mathbf{V}_j = \overline{(\bigcup_{j=-\infty}^{+\infty} \mathbf{V}_j)} = \mathbf{L}^2(\mathbb{R}) \tag{4.5}$$

$\bullet$ *There exists $\theta(\cdot)$ such that $\{\theta(t - n)\}_{n \in \mathbb{Z}}$ is a Riesz basis of $\mathbf{V}_0$*

Example: Shannon approximations Frequency band-limited functions yield a multiresolution approximations. The space $\mathbf{V}_j$ is defined as the set of functions whose Fourier transform has a support included in $[-2^{-j}\pi, 2^{-j}\pi]$. It can be shown that an orthonormal basis $\{\theta(t - n)\}_{n \in \mathbb{Z}}$ of $\mathbf{V}_0$ is defined by

$$\theta(t) = \frac{\sin(\pi t)}{\pi t}. \tag{4.6}$$

All the other properties of the multiresolution approximation are easily verified.

The approximation of f at the resolution of 2^{-j} is defined as the orthogonal projection $P_{\mathbf{V}_j} f$ on $\mathbf{V}_j$. To compute this projection, we must find an orthonormal basis of $\mathbf{V}_j$. The following theorem orthogonalizes the Riesz basis $\{\theta(t - n)\}_{n \in \mathbb{Z}}$ and constructs an orthonormal basis of each subspace $\mathbf{V}_j$ by dilating and translating a single function $\phi(\cdot)$ called a *scaling function*. To avoid the resolution 2^{-j} and the scale 2^j, in the rest of the chapter the notation of resolution is dropped and $P_{\mathbf{V}_j} f$ is called an approximation at

the scale 2^j.

Theorem 4.1 *Let* $\{V_j\}_{j \in \mathbb{Z}}$ *be a multiresolution approximation and* ϕ *be the scaling function whose Fourier transform is*

$$\hat{\phi}(\omega) = \frac{\hat{\theta}(\omega)}{(\sum_{k=-\infty}^{+\infty} |\hat{\theta}(\omega + 2k\pi)|^2)^{1/2}} \tag{4.7}$$

Let us denote

$$\phi_{j,n}(t) = \frac{1}{\sqrt{2^j}} \phi(\frac{t - 2^j n}{2^j}). \tag{4.8}$$

The family $\{\phi_{j,n}\}_{n \in \mathbb{Z}}$ *is an orthonormal basis of* V_j *for all* $j \in \mathbb{Z}$.

The orthogonal projection of f over V_j is obtained with an expression in the scaling orthogonal basis

$$P_{V_j} f = \sum_{k=-\infty}^{+\infty} < f, \phi_{j,n} > \phi_{j,n}. \tag{4.9}$$

The inner products

$$a_j[n] = < f, \phi_{j,n} > \tag{4.10}$$

provide a discrete approximation of f at the scale 2^j.

For the case of Shannon approximations, we have constructed Riesz basis $\{\theta(t-n)\}_{n \in \mathbb{Z}}$, which are orthonormal basis, hence $\phi(t) = \theta(t)$.

A multiresolution approximation is entirely characterized by the scaling function ϕ that generates an orthonormal basis of each subspace V_j. We study the properties of ϕ which guarantee that the spaces V_j satisfy all the conditions of a multiresolution approximation. It is proven [67] that any scaling function can be specified by a discrete filter called a conjugate mirror filter [69]. The multiresolution causality property imposes that $V_j \subset V_{j-1}$. In particular, $1/\sqrt{2}\phi(t/2) \in V_1 \subset V_0$. Since $\{\phi(t-n)\}_{n \in \mathbb{Z}}$ is a basis of V_0, we can decompose

$$\frac{1}{\sqrt{2}}\phi(\frac{t}{2}) = \sum_{n=-\infty}^{+\infty} h[n]\phi(t - n), \tag{4.11}$$

with

$$h[n] = < \frac{1}{\sqrt{2}}\phi(\frac{t}{2}), \phi(t - n) > .$$ (4.12)

Equation (4.11) relates a dilation of the scaling function ϕ by 2 to its integer translations. The sequence $h[n]$ can be interpreted as a discrete filter.

The Fourier transformation of both sides of Equation (4.11) yields

$$\hat{\phi}(2\omega) = \frac{1}{\sqrt{2}}\hat{h}(\omega)\hat{\phi}(\omega)$$ (4.13)

for $\hat{h}(\omega) = \sum_{n=-\infty}^{+\infty} h[n]e^{-j\omega n}$. It then makes sense to express $\hat{\phi}(\omega)$ directly as a product of dilation of $\hat{h}(\omega)$. For any $p \geq 0$, Equation (4.13) implies

$$\hat{\phi}(2^{-p+1}\omega) = \frac{1}{\sqrt{2}}\hat{h}(2^{-p}\omega)\hat{\phi}(2^{-p}\omega).$$ (4.14)

If $\hat{\phi}(\omega)$ is continuous at $\omega = 0$ then $\lim_{p->+\infty}\hat{\phi}(2^{-p}\omega) = \hat{\phi}(0)$. By substitution, we obtain

$$\hat{\phi}(\omega) = \prod_{p=1}^{+\infty} \frac{\hat{h}(2^{-p}\omega)}{\sqrt{2}}\hat{\phi}(0).$$ (4.15)

The following theorem by Mallat and Meyer [67] gives necessary and sufficient conditions on $\hat{h}(\omega)$ to guarantee that this infinite product is the Fourier transform of a scaling function.

Theorem 4.2 *Let $\phi \in \mathbf{L}^2(\mathbb{R})$ be an integrable scaling function. The Fourier transform of $h[n] = < \frac{1}{\sqrt{2}}\phi(\frac{t}{2}), \phi(t - n) >$ satisfies*

$$\forall \omega \in \mathbb{R}, \ |\hat{h}(\omega)|^2 + |\hat{h}(\omega + \pi)|^2 = 2,$$ (4.16)

and

$$\hat{h}(0) = \sqrt{(2)}.$$ (4.17)

Conversely, if $\hat{h}(\omega)$ is 2π periodic and continuously differentiable in a neighborhood of $\omega = 0$, if it satisfies Equation (4.16) and (4.17), then

$$\hat{\phi}(\omega) = \prod_{p=1}^{\infty} \frac{\hat{h}(2^{-p}\omega)}{\sqrt{2}}.$$ (4.18)

is the Fourier transform of a scaling function $\phi \in \mathbf{L}^2(\mathbb{R})$.

For the case of Shannon multiresolution approximation, $\hat{\phi}(\omega) = \mathbf{1}_{[-\pi,\pi]}(\omega)$. We thus derive from Equation (4.18) that

$$\forall \omega \in [-\pi, \pi], \ \hat{h}(\omega) = \sqrt{2}\mathbf{1}_{[-\pi/2,\pi/2]}(\omega).$$

4.2 Orthonormal Wavelet Basis

In the multiscale analysis, we are especially interested in the difference between consecutive resolution scales. This difference is often called a detail signal. The approximation at resolution 2^{j-1} and 2^j of a signal are respectively equal to their orthogonal projections on $\mathbf{V}_{j-1}$ and $\mathbf{V}_j$. It can also be shown that the signal details at resolution 2^j are given by an orthogonal projection of the original signal onto the orthogonal complement of $\mathbf{V}_j$ in $\mathbf{V}_{j-1}$. Let $\mathbf{O}_j$ represent the orthogonal complement of $\mathbf{V}_j$, we then have,

$$\mathbf{O}_j \bot \mathbf{V}_j,$$

$$\mathbf{O}_j \oplus \mathbf{V}_j = \mathbf{V}_{j-1},$$

where $\oplus$ denotes a direct sum. In order to obtain the detail signal of a function on $\mathbf{O}_j$, we need to find an orthonormal basis of $\mathbf{O}_j$.

Theorem 4.3 *[70] Let ϕ be a scaling function and h the corresponding conjugate mirror filter. Let $\psi(t)$ be function whose Fourier transformation is defined by*

$$\hat{\psi}(\omega) = \frac{1}{\sqrt{2}}\hat{g}(\frac{\omega}{2})\hat{\phi}(\frac{\omega}{2}) \tag{4.19}$$

with

$$\hat{g}(\omega) = e^{-j\omega}\hat{h}^*(\omega + \pi) \tag{4.20}$$

Let us denote

$$\psi_{j,n}(t) = \frac{1}{\sqrt{2^j}}\psi(\frac{t - 2^j n}{2^j}) \tag{4.21}$$

then, for any scale 2^j, $\{\psi_{j,n}\}_{n\in\mathbb{Z}}$ is an orthonormal basis of $\mathbf{O}_j$. For all scales, $\{\psi_{j,n}\}_{(j,n)\in\mathbb{Z}^2}$ is an orthonormal basis of $\mathbf{L}^2(\mathbb{R})$.

The orthogonal projection of a signal $f \in \mathbf{L}^2(\mathbb{R})$ in a detailed space $\mathbf{O}_j$ is then obtained with a partial expansion in its wavelet basis

$$P_{\mathbf{O}_j} f = \sum_{n=-\infty}^{+\infty} < f, \psi_{j,n} > \psi_{j,n}. \tag{4.22}$$

A signal expansion in a wavelet orthonormal basis can thus be viewed as an aggregation of details at all the scales 2^j that goes from $-\infty$ to $+\infty$,

$$f = \sum_{j=-\infty}^{+\infty} P_{\mathbf{O}_j} f = \sum_{j=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} < f, \psi_{j,n} > \psi_{j,n}. \tag{4.23}$$

Many applications using wavelet decomposition desire efficient approximations of particular classes of functions by a few non-zero coefficients. This usually requires optimizing the design of $\psi(\cdot)$ to produce maximum number of wavelet coefficients $< f, \psi_{j,n} >$ that are close to zero. The actual number of coefficients with non-negligible values depends on the regularity of f, the number of vanishing moments of the analyzing wavelet and the size of its support.

A wavelet has p vanishing moments [70] if

$$\int_{-\infty}^{+\infty} t^k \psi(t) dt = 0, \quad 0 \leq k \leq p \tag{4.24}$$

The vanishing moment is crucial to measure the local regularity of a signal. If the wavelet has p vanishing moments, then it can be shown that the wavelet transformation is actually a multiscale differential operator of order p. This nice property relates the differentiability of f with its wavelet transform decay at fine scales.

The kth order derivative $\hat{\psi}^{(k)}(\omega)$ is the Fourier transform of $(-it)^k \psi(t)$. Hence

$$\hat{\psi}^{(k)}(0) = \int_{-\infty}^{+\infty} (-it)^k \psi(t) dt. \tag{4.25}$$

If a wavelet $\psi(\cdot)$ has p vanishing moments, then by Equation (4.24), $\hat{\psi}^{(k)}(\omega)$ and its first $p-1$ derivatives are zero at $\omega = 0$. Theory (4.3) shows that

$$\sqrt{2}\hat{\psi}(2\omega) = e^{-i\omega} \hat{h}^*(\omega + \pi) \hat{\phi}(\omega). \tag{4.26}$$
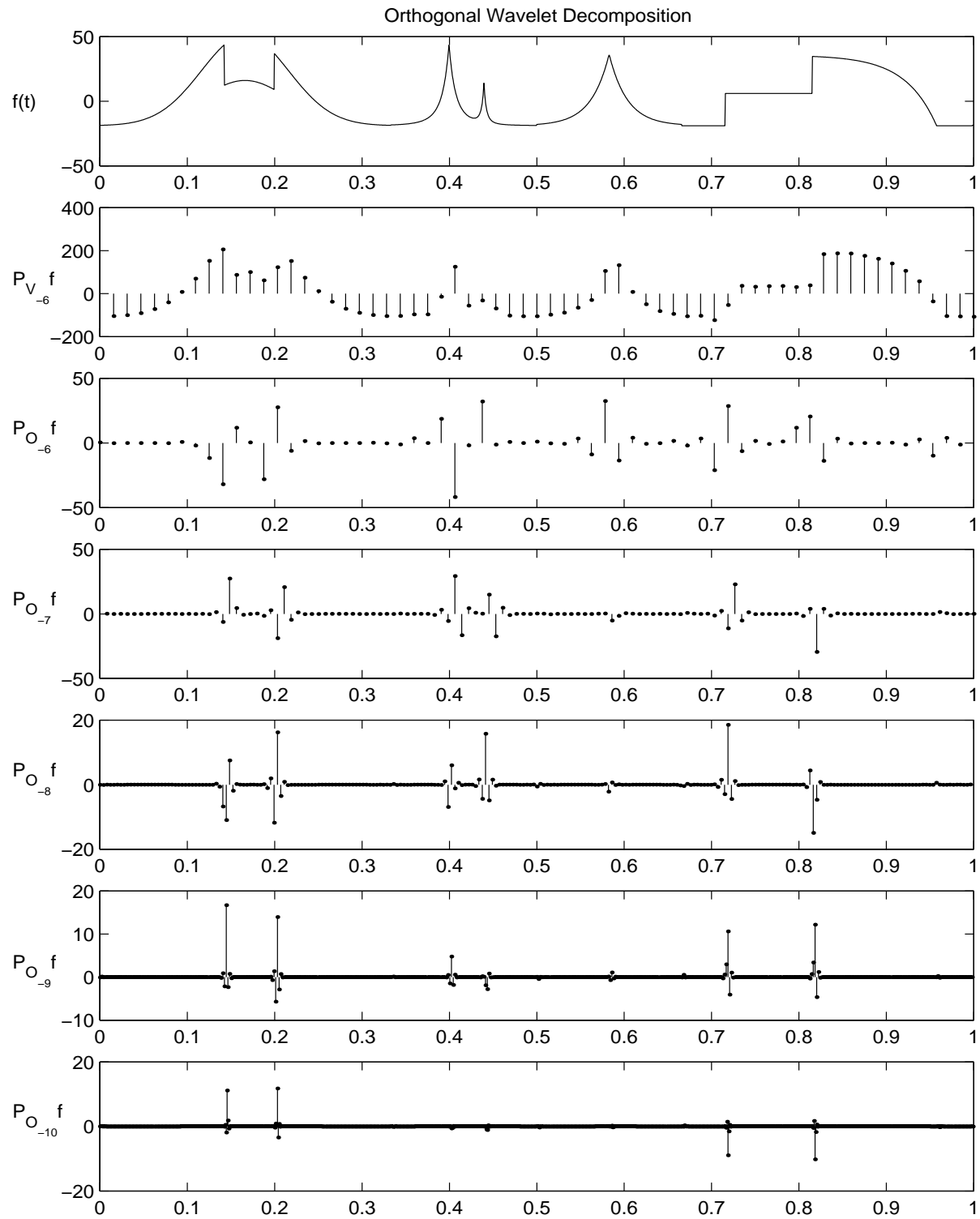
Figure 4.1: Orthogonal wavelet decomposition of $f(t)$ with Daubechies wavelet at resolutions of 2^j, $j = -6, -7, ..., -10$, $P_{\mathbf{V}_{-6}}f$ is the remaining coarse signal approximation $a_J[n] = <f, \phi_{J,n}>$ for $J = -6$.

Since $\hat{\phi}(0) \neq 0$, it follows that if a wavelet $\psi(\cdot)$ has p vanishing moments, then its corresponding conjugate mirror filter $\hat{h}(\omega)$ and its first $p - 1$ derivatives are zero at $\omega = \pi$, i.e., we can decompose $\hat{h}(\omega)$ as

$$\hat{h}(\omega) = \sqrt{2} \left(\frac{1 + e^{-i\omega}}{2}\right)^p R(\omega) \tag{4.27}$$

where $R(\omega)$ is a function of $e^{-i\omega}$.

As mentioned before, the number of wavelet coefficients of f with non-negligible values depends on not only the number of vanishing moment, but also the size of its support. Suppose $f(t)$ has a singularity point at $t = t_0$, and if t_0 is inside the support of $\psi_{j,n}(t) = \sqrt{2^{-j}}\psi(2^{-j}t - n), \forall (j, n) \in \mathbb{Z}^2$, then the corresponding wavelet coefficient

$$< f, \psi_{j,k} >= \int_{-\infty}^{+\infty} f(t)\frac{1}{\sqrt{2^j}}\psi(\frac{t - 2^j n}{2^j})dt$$

might have a large value. If $\psi_{j,n}$ has a compact support of size N, then at scale 2^j, there must be N wavelets $\psi_{j,n}$ whose support includes the singularity point at t_0. To minimize the number of coefficients with non-negligible values, we have to choose a wavelet $\psi(\cdot)$ with a small support size.

The support size of $\psi(\cdot)$ are related to the support size of $h(\cdot)$ and $\phi(\cdot)$. Daubechies [29] showed that the scaling function $\phi(\cdot)$ has a compact support if and only if $h(\cdot)$ has a compact support. Assume the support of $h(\cdot)$ and $\phi(\cdot)$ are $[N_1, N_2]$ and $[K_1, K_2]$ respectively, recall that

$$\frac{1}{\sqrt{2}}\phi(\frac{t}{2}) = \sum_{n=-\infty}^{+\infty} h[n]\phi(t - n),$$

the support size of $\phi(t/2)$ should be $[N_1 + K_1, N_2 + K_2]$. On the other hand, $\phi(t/2)$ is just a dilation of $\phi(t)$, its support should be $[2K_1, 2K_2]$. The equality proves that if h has a compact support of $[N_1, N_2]$, then the support of $\phi(t)$ is also $[N_1, N_2]$.

Since $\psi(t/2) \in \mathbf{O}_1 \subset \mathbf{V}_0$, it thus can be decomposed in $\{\phi(t - n)\}_{n \in \mathbb{Z}}$, which is an orthonormal basis of $\mathbf{V}_0$,

$$\frac{1}{\sqrt{2}}\psi(\frac{t}{2}) = \sum_{n=-\infty}^{+\infty} g[n]\phi(t - n).$$

	n	h[n]		n	h[n]		n	h[n]
$p=2$	0	.482962913145	$p=2$	0	.332670552950	$p=4$	0	.230377813309
	1	.836516303738		1	.806891509311		1	.714846570553
	2	.224143868042		2	.459877502118		2	.630880767930
	3	-.129409522551		3	-.135011020010		3	-.027983769417
				4	-.085441273882		4	-.187034811719
				5	.035226291882		5	.030841381836
							6	.032883011667
							7	-.010597401785

Table 4.1: Daubechies filter coefficients for wavelets with $p = 2, 3, 4$ vanishing moments.

By equation (4.20), we have

$$g[n] = (-1)^{1-n} h[1-n].$$

Then we obtain that

$$\frac{1}{\sqrt{2}} \psi(\frac{t}{2}) = \sum_{n=-\infty}^{+\infty} (-1)^{1-n} h[1-n] \phi(t-n).$$

The sum in the right hand side has a support equal to $[N_1 - N_2 + 1, N_2 - N_1 + 1]$. Hence the support of $\psi(t)$ is $[(N_1 - N_2 + 1)/2, \ (N_2 - N_1 + 1)/2]$.

The direct consequence of the above derivation is that tradeoff should be made between the support size of a wavelet and its number of vanishing moment. One can not have a wavelet of compact support with an arbitrary high vanishing moments. To ensure a wavelet ψ with p vanishing moments has a minimal support, we need construct $\hat{h}(\omega)$ as in Equation (4.27) to have a minimum degree. The difficulty is to design a polynomial $R(\omega)$ of minimum degree such that

$$|\hat{h}(\omega)|^2 + |\hat{h}(\omega + \pi)|^2 = 2. \tag{4.28}$$

Since $h[n]$ is real, $|\hat{h}(\omega)|^2$ is a even function and thus can be written as a polynomial in $\cos \omega$. Hence $|R(\omega)|^2$ defined in Equation (4.27) is a polynomial in $\cos \omega$ and we can also
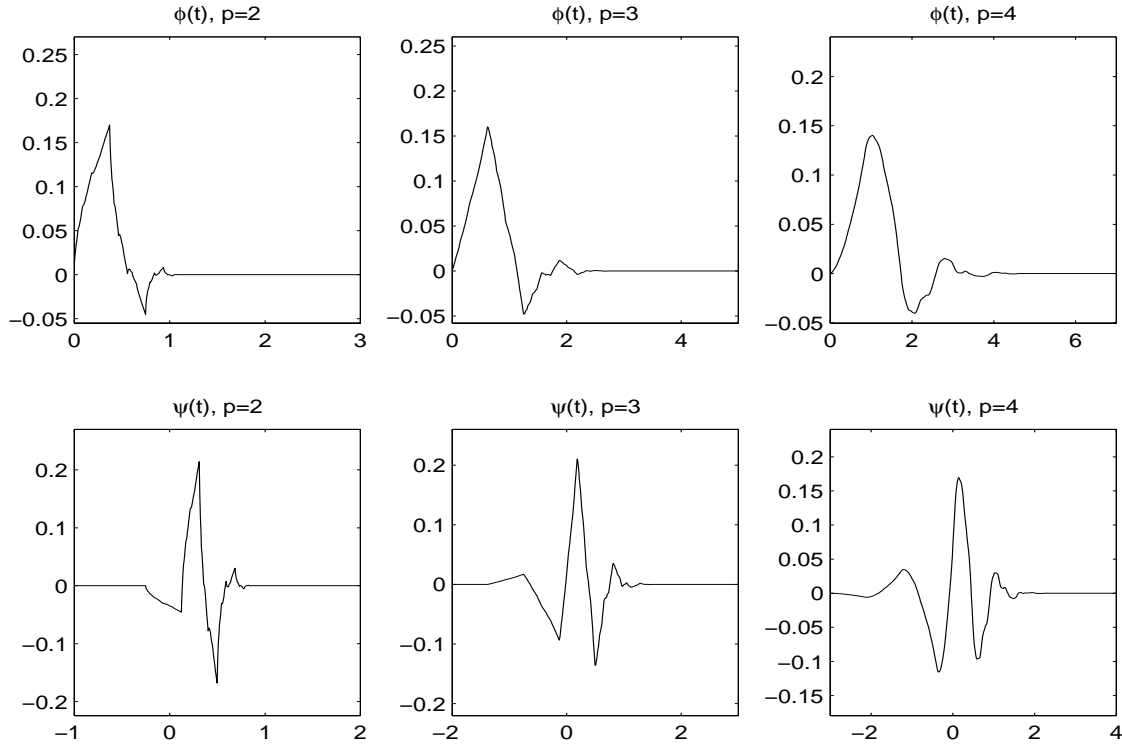
Figure 4.2: Daubechies scaling function and wavelets

write it as a polynomial $P(\sin^2 \omega/2)$

$$|\hat{h}(\omega)|^2 = 2(\cos^2 \frac{\omega}{2})^p P(\sin^2 \frac{\omega}{2}).$$

The quadrature condition (4.28) is equivalent to

$$(1-y)^p P(y) + y^p P(1-y) = 1, \tag{4.29}$$

for any $y = \sin^2(\omega/2) \in [0,1]$. To minimize the nonzero terms of the finite Fourier series $\hat{h}(\omega)$, we must find the solution $P(y) \geq 0$ of minimum degree, which is obtained with the Bezout [71] classical theorem on polynomials. The Bezout theorem proves that there exist two unique polynomials $P_1(y)$ and $P_2(y)$ such that

$$(1-y)^p P_1(y) + y^p P_2(y) = 1.$$

It can be verified that $P(y) = P_1(y) = P_2(1-y)$ with

$$P(y) = \sum_{k=0}^{p-1} \binom{p-1+k}{k} y^k.$$
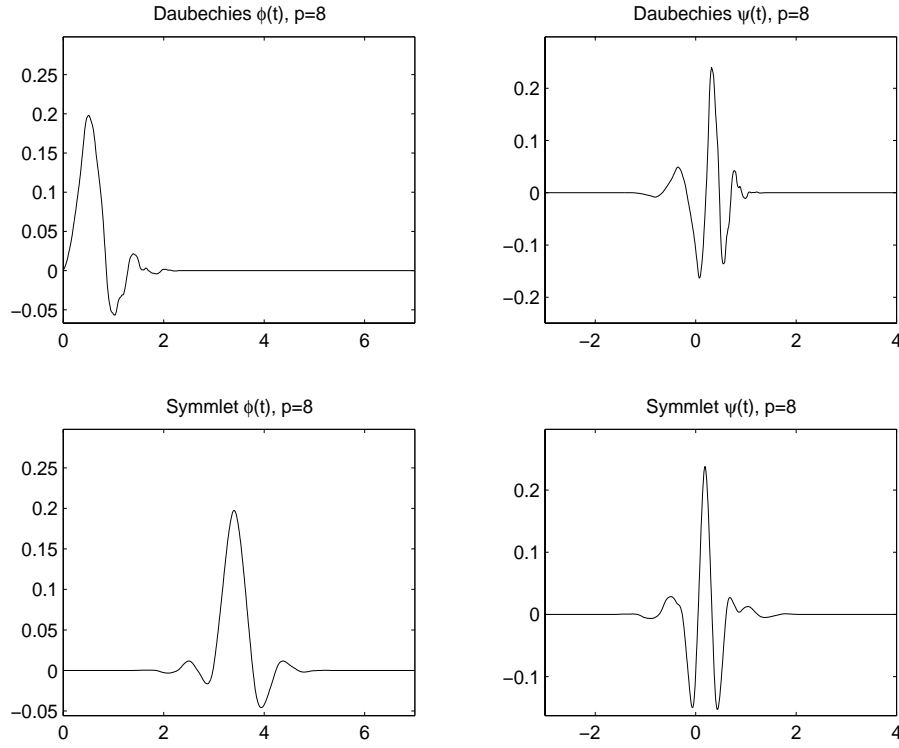
Figure 4.3: Symmlet scaling functions and wavelets

Clearly $P(y) \geq 0$ for $y = \sin^2(\omega/2) \in [0, 1]$. Hence $P(y)$ is the polynomial of the minimum degree satisfying the Equation (4.29).

Now we need to construct a minimum degree polynomial

$$R(e^{i\omega}) = \sum_{k=0}^{m} r_k e^{-ik\omega} = r_0 \prod_{k=0}^{m} (1 - a_k e^{-i\omega})$$

such that $|R(e^{i\omega})|^2 = P(\sin^2 \omega/2)$. Let $z = e^{-i\omega}$, we obtain

$$R(z)R(z^{-1}) = r_0^2 \prod_{k=0}^{m} (1 - a_k z)(1 - a_k z^{-1}) = P(\frac{2 - z - z^{-1}}{4}) = Q(z) \qquad (4.30)$$

Since $Q(z) = R(z)R(z^{-1})$ has real coefficients, if c_k is a root of $Q(z)$, then c_k^*, $1/c_k$ and $1/c_k^*$ are roots of $Q(z)$ as well. To design $R(z)$ satisfies Equation (4.30), we choose each root a_k of $R(z)$ among a pair $(c_k, 1/c_k)$ and include a_k^* to obtain real coefficients. This procedure yields a polynomial of minimum degree $m = p - 1$. the resulting filter h of minimum size has $N = p + m + 1 = 2p$ nonzero coefficients. Among all the

possible factorizations, the minimum phase solution $R(e^{i\omega})$ is obtained by choosing a_k among $(c_k, 1/c_k)$ to be inside the unit circle $|a_k| \leq 1$. The resulting causal filter h is a Daubechies filter of order p. Table (4.1) gives the coefficients of these Daubechies filters for $p = 2, 3, 4$.

If ψ is a wavelet with p vanishing moments that generates an orthonormal basis of $\mathbf{L}^2(\mathbb{R})$, then it has a support size larger or equal to $2p-1$. A Daubechies wavelet [29] calculated with Daubechies filter banks has a minimum size of support equal to $[-p+1, p]$. The support of the corresponding scaling functions ϕ is $[0, 2p - 1]$.

As a example, Figure (4.2) shows Daubechies scaling functions ϕ and wavelet ψ at scale $j = 3$ with $p = 2, 3, 4$ vanishing moments respectively.

Daubechies wavelets are very asymmetric because they are constructed by selecting the minimum phase square root of $Q(z)$ in Equation (4.30). One can show that filters corresponding to a minimum phase square root have their energy optimally concentrated near the starting point of their support. They are thus highly non-symmetric, which yields very asymmetric wavelets.

To obtain symmetric or antisymmetric wavelet, the filter h must be symmetric or antisymmetric with respect to the center of its support, which means that $\hat{h}(\omega)$ has a linear phase. Haar filter is the only real compactly supported conjugate mirror filter that has a linear phase. The *Symmlet* filters of Daubechies are obtained by optimizing the choice of the square root of $Q(z)$ to obtain an almost linear phase. The resulting wavelet still have a minimum support $[-p + 1, p]$ with p vanishing moments but they are more symmetric, as illustrated in Figure(4.3).

4.3   Fast Orthogonal Wavelet Transform

A fast wavelet transform decomposes successively each approximation $P_{\mathbf{V}_j}$ into a coarser resolution $P_{\mathbf{V}_{j+1}}$ plus the wavelet coefficients carried by $P_{\mathbf{O}_{j+1}}$. In the reverse direction, the reconstruction from wavelet coefficients recovers each $P_{\mathbf{V}_j}$ from $P_{\mathbf{V}_{j+1}}$ and $P_{\mathbf{O}_{j+1}}$.

Since $\{\phi_{j,n}\}_{n\in\mathbb{Z}}$ and $\{\psi_{j,n}\}_{n\in\mathbb{Z}}$ are orthonormal basis of $\mathbf{V}_j$ and $\mathbf{O}_j$, the projection of $f \in \mathbf{L}^2(\mathbb{R})$ on $\mathbf{V}_J$ is characterized by

$$P_{\mathbf{V}_J}f = \sum_{j=J+1}^{L}\sum_{n=-\infty}^{+\infty} d_j[n]\psi_{j,n} + \sum_{n=-\infty}^{+\infty} a_L[n]\phi_{L,n}, \forall L > J \tag{4.31}$$

where

$$a_j[n] = <f, \phi_{j,n}>, (j,n) \in \mathbb{Z}^2 \tag{4.32}$$

and

$$d_j[n] = <f, \psi_{j,n}>, (j,n) \in \mathbb{Z}^2. \tag{4.33}$$

The following theorem [67] shows that these coefficients are calculated with a cascade of discrete convolution and subsampling. We denote $\bar{x}[n] = x[-n]$ and a zero interpolation of $x[n]$ as

$$\check{x}[n] = \begin{cases} x[k], & n = 2k \\ 0, & n = 2k+1 \end{cases}.$$

Theorem 4.4 *Let ϕ be a scaling function and $\{h, g\}$ are its corresponding conjugate mirror filters. At the decomposition,*

$$a_{j+1}[k] = \sum_{n=-\infty}^{+\infty} h[n-2k]a_j[n] = a_j \star \bar{h}[2k] \tag{4.35}$$

$$d_{j+1}[k] = \sum_{n=-\infty}^{+\infty} g[n-2k]a_j[n] = a_j \star \bar{g}[2k] \tag{4.36}$$

At the reconstruction,

$$a_j[k] = \sum_{n=-\infty}^{+\infty} h[k-2n]a_{j+1}[n] + \sum_{n=-\infty}^{+\infty} g[k-2n]d_{j+1}[n] = \check{a}_{j+1} \star h[k] + \check{d}_{j+1} \star g[k]. \tag{4.37}$$

Theory (4.4) proves that a_{j+1} and d_{j+1} are computed by taking every other sample of the convolution of a_j with $\bar{h}$ and $\bar{g}$ respectively, as illustrated by Figure (4.4a). The filter $\bar{h}$ removes the higher frequencies of the inner product sequence a_j whereas $\bar{g}$ is a high-pass filter which collects the remaining highest frequencies. The reconstruction is an interpolation that inserts zeros to expand a_{j+1} and d_{j+1} and filter these signals, as shown in Figure (4.4b).

(a)

(b)
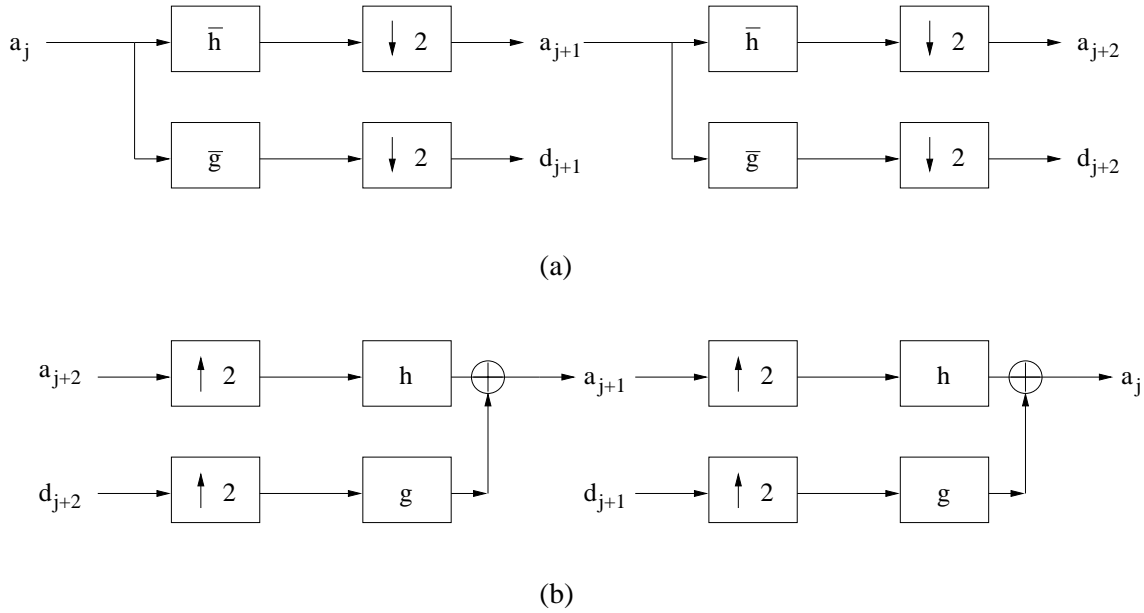
Figure 4.4: (a) A fast wavelet transform is computed with a cascade of filters with $\bar{h}$ and $\bar{g}$ followed by a factor 2 subsampling. (b) A fast inverse wavelet transform reconstructs progressively each a_j by inserting zeros between samples of a_{j+1} and d_{j+1}, filtering and adding the outputs.

The decomposition of a discrete signal in conjugate mirror filters $\{h, g\}$ can be interpreted as an expansion in a basis of $\mathbf{l}^2(\mathbb{Z})$, the resulting family $\{h[n - 2l], g[n - 2l]\}_{l \in \mathbb{Z}}$ is an orthogonal basis of $\mathbf{l}^2(\mathbb{Z})$. An orthogonal wavelet representation of $a_J =< f, \phi_{J,n} >$ is composed of wavelet coefficients of f at scales $2^J < 2^j \leq 2^L$ plus the remaining approximation at the largest scale 2^L,

$$\left[\ \{d_j\}_{J < j \leq L}, a_L\ \right]$$

It is computed by iterating Equation (4.35) and (4.36) for $J < j \leq L$. Figure (4.1) gives a numerical example computed with the Daubechies filter of order $p = 4$. The original signal a_J is recovered from this wavelet representation by iterating the reconstruction Equation (4.37) for $J < j \leq L$.

The conjugate mirror filters are often used in filter banks that cascade several levels of filtering and subsampling. It is thus necessary to understand the behavior of such a cascade [72]. In a wavelet filter bank tree, the output of the low-pass filter $\bar{h}$ is sub-

decomposed whereas the output of the high-pass filter $\bar{g}$ is not. This is illustrated in Figure (4.4). Suppose that the sampling distance of the original signal is N^{-1}. We denote $a_J[n]$ this discrete signal, with $2^J = N^{-1}$. At the depth $j > J$ of this filter bank tree, the low-pass signal a_j and high-pass signal d_j can be written as

$$a_j[n] = a_J \star \bar{\phi}_j[2^{j-J}n] = < a_J[k], \phi_j[k - 2^{j-J}n] >$$

and

$$d_j[n] = a_J \star \bar{\psi}_j[2^{j-J}n] = < a_J[k], \psi_j[k - 2^{j-J}n] > .$$

The Fourier transform of these equivalent filters are

$$\hat{\phi}(\omega) = \prod_{p=0}^{j-J-1} \hat{h}(2^p\omega) \tag{4.38}$$

and

$$\hat{\psi}(\omega) = \hat{g}(2^{j-J-1}\omega) \prod_{p=0}^{j-J-2} \hat{h}(2^p\omega). \tag{4.39}$$

A filter bank tree of depth $L - J > 0$ decomposes a_J over the family of vectors

$$\left[\{\phi_L[k - 2^{L-J}n]\}_{n\in\mathbb{Z}}, \; \{\psi_j[k - 2^{j-J}n]\}_{J<j\leq L,n\in\mathbb{Z}} \right]. \tag{4.40}$$

For conjugate mirror filters, one can verify that this family is an orthonormal basis of $\mathbf{I}^2(\mathbb{Z})$. These discrete vectors are close to a uniform sampling of the continuous time scaling functions $\phi_j(t) = 2^{-j/2}\phi(2^{-j}t)$ and wavelets $\psi_j(t) = 2^{-j/2}\psi(2^{-j}t)$. When the number $L - J$ of successive convolutions increases, one can verify that $\phi_j[n]$ and $\psi_j[n]$ converges respectively to $N^{-1/2}\phi_j(N^{-1}n)$ and $N^{-1/2}\psi_j(N^{-1}n)$. We therefore refer the Equation (4.40) to as a **Discrete Wavelet Basis** of $\mathbf{I}^2(\mathbb{Z})$.

4.4 Filter Banks and Biorthogonal Wavelets

The fast discrete wavelet transform decomposes signal into low-pass and high-pass components subsampled by 2, the inverse transform performs the reconstruction. Study of such classical multirate filter banks became a major signal processing topic in 1976,
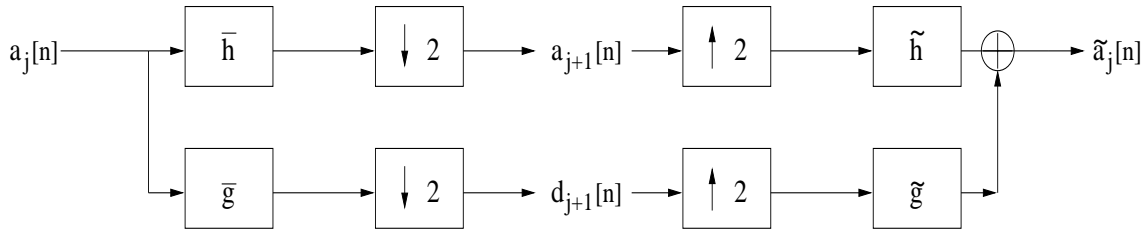
Figure 4.5: A two channel perfect reconstruction filter banks

when Croisier, Esteban and Galand [73] discovered that it is possible to perform such decompositions and reconstructions with *quadrature mirror filters*. However, besides the simple Haar filter, a quadrature mirror filter can not have a finite impulse response. In 1984, Smith and Barnwell [74] and Mintzer [75] found necessary and sufficient conditions for obtaining perfect reconstruction orthogonal filters with a finite impulse response, that they called *conjugate mirror filters*. The theory was completed by the *biorthogonal mirror filter* equations of Vetterli [76, 77]. We follow this digital signal processing approach which gives a simple understanding of perfect reconstruction filter banks.

4.4.1 Perfect Reconstruction Filter Banks

A two-channel multirate filter bank convolves a signal a_0 with a low-pass filter $\bar{h}[n] = h[n]$ and a high-pass filter $\bar{g}[n] = g[n]$ and subsamples the output by 2,

$$a_1[n] = a_0 \star \bar{h}[2n] \ \text{ and } \ d_1[n] = a_0 \star \bar{g}[2n]. \tag{4.41}$$

A reconstructed signal $\tilde{a}_0$ is obtained by filtering the zero interpolated signals with a dual low-pass filter $\tilde{h}$ and a dual high-pass filter $\tilde{g}$, as shown in Figure (4.5). With the zero insertion notation (4.34), it yields

$$\tilde{a}_0[n] = \breve{a}_1 \star \tilde{h}[n] + \breve{d}_1 \star \tilde{g}[n]. \tag{4.42}$$

The following theorem by Vetterli [77] gives biorthogonal conditions, which guarantee that $a_0 = \tilde{a}_0$.

Theorem 4.5 *The filter bank performs an exact reconstruction for any input signal if and only if*

$$\hat{h}^*(\omega + \pi)\hat{\tilde{h}}(\omega) + \hat{g}^*(\omega + \pi)\hat{\tilde{g}}(\omega) = 0, \tag{4.43}$$

and

$$\hat{h}^*(\omega)\hat{\tilde{h}}(\omega) + \hat{g}^*(\omega)\hat{\tilde{g}}(\omega) = 2. \tag{4.44}$$

Theory (4.5) proves that the reconstruction filters $\tilde{h}$ and $\tilde{g}$ are entirely specified by the decomposition filters h and g. In the matrix form, the biorthogonal condition can be written as,

$$\begin{bmatrix} \hat{h}(\omega) & \hat{g}(\omega) \\ \hat{h}(\omega + \pi) & \hat{g}(\omega + \pi) \end{bmatrix} \times \begin{bmatrix} \hat{\tilde{h}}^*(\omega) \\ \hat{\tilde{g}}^*(\omega) \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}.$$

For the finite impulse response filters, there exist $a \in \mathbb{R}$ and $l \in \mathbb{Z}$ such that

$$\hat{g}(\omega) = ae^{-i(2l+1)\omega}\hat{\tilde{h}}^*(\omega + \pi) \text{ and } \hat{\tilde{g}}(\omega) = a^{-1}e^{-i(2l+1)\omega}\hat{h}^*(\omega + \pi).$$

The factor a is a gain which is inverse for the decomposition and reconstruction filters and l is a reverse shift. We generally set $a = 1$ and $l = 0$. The internal relationship between biorthogonal filter banks become

$$\hat{g}(\omega) = e^{-i\omega}\hat{\tilde{h}}^*(\omega + \pi) \text{ and } \hat{\tilde{g}}(\omega) = e^{-i\omega}\hat{h}^*(\omega + \pi). \tag{4.45}$$

In the time domain, Equation (4.45) can be written as

$$g[n] = (-1)^{1-n}\tilde{h}[1 - n] \text{ and } \tilde{g}[n] = (-1)^{1-n}h[1 - n]. \tag{4.46}$$

The two pairs of filters $\{h, g\}$ and $\{\tilde{h}, \tilde{g}\}$ plays a symmetric role and can be inverted. The biorthogonal condition simplifies to

$$\hat{h}^*(\omega)\hat{\tilde{h}}(\omega) + \hat{h}^*(\omega + \pi)\hat{\tilde{h}}(\omega + \pi) = 2. \tag{4.47}$$

If we impose that the decomposition filter h is equal to the reconstruction filter $\tilde{h}$, then Equation (4.47) is exactly the condition of conjugate mirror filters given in Equation (4.28).

4.4.2   Biorthogonal Wavelets

An infinite cascade of perfect reconstruction filters $\{h, g\}$ and $\{\tilde{h}, \tilde{g}\}$ yields two scaling functions and wavelets whose Fourier transforms satisfy

$$\hat{\phi}(\omega) = \prod_{p=1}^{+\infty} \frac{\hat{h}(2^{-p}\omega)}{\sqrt{2}} \;\; \text{and} \;\; \hat{\tilde{\phi}}(\omega) = \prod_{p=1}^{+\infty} \frac{\hat{\tilde{h}}(2^{-p}\omega)}{\sqrt{2}} \tag{4.48}$$

and

$$\hat{\psi}(\omega) = \frac{1}{\sqrt{2}}\hat{g}(\frac{\omega}{2})\hat{\phi}(\frac{\omega}{2}) \;\; \text{and} \;\; \hat{\tilde{\psi}}(\omega) = \frac{1}{\sqrt{2}}\hat{\tilde{g}}(\frac{\omega}{2})\hat{\tilde{\phi}}(\frac{\omega}{2}). \tag{4.49}$$

One [71, 78] can proves, with some additional conditions, $\hat{\phi}$ and $\hat{\tilde{\phi}}$ are the Fourier transforms of finite energy functions. The two wavelet families $\{\psi_{j,n}\}_{(j,n)\in\mathbb{Z}^2}$ and $\{\tilde{\psi}_{j,n}\}_{(j,n)\in\mathbb{Z}^2}$ are biorthogonal Riesz bases of $\mathbf{L}^2(\mathbb{R})$. Biorthogonality means that for any $(j_1, j_2, n_1, n_2) \in \mathbb{Z}^4$,

$$< \psi_{j_1,n_1}, \tilde{\psi}_{j_2,n_2} >= \delta(j_1 - j_2, n_1 - n_2) \tag{4.50}$$

Any $f \in \mathbf{L}^2(\mathbb{R})$ has two possible decompositions in these bases,

$$f = \sum_{j=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} < f, \psi_{j,n} > \tilde{\psi}_{j,n} = \sum_{j=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} < f, \tilde{\psi}_{j,n} > \psi_{j,n}. \tag{4.51}$$

The Riesz stability implies that there exist $A > 0$ and $B > A$ such that

$$A\|f\|^2 \leq \sum_{j=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} | < f, \psi_{j,n} > |^2 \leq B\|f\|^2$$

and

$$\frac{1}{B}\|f\|^2 \leq \sum_{j=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} | < f, \tilde{\psi}_{j,n} > |^2 \leq \frac{1}{A}\|f\|^2.$$

Biorthogonal wavelet bases are related to multiresolution approximations. The family $\{\phi(t-n)\}_{n\in\mathbb{Z}}$ is a Riesz basis of the space $\mathbf{V}_0$ it generates, whereas $\{\tilde{\phi}(t-n)\}_{n\in\mathbb{Z}}$ is a Riesz basis of another space $\tilde{\mathbf{V}}_0$. Let $\mathbf{V}_j$ and $\tilde{\mathbf{V}}_j$ be the spaces spanned by $\{\phi_{j,n}(t)\}_{n\in\mathbb{Z}}$ and $\{\tilde{\phi}_{j,n}(t)\}_{n\in\mathbb{Z}}$ respectively. One can verify that $\{\mathbf{V}_j\}_{n\in\mathbb{Z}}$ and $\{\tilde{\mathbf{V}}_j\}_{n\in\mathbb{Z}}$ are two multiresolution approximations of $\mathbf{L}^2(\mathbb{R})$. For any $j \in \mathbb{Z}$, the dilated $\{\psi_{j,n}(t)\}_{n\in\mathbb{Z}}$ and $\{\tilde{\psi}_{j,n}(t)\}_{n\in\mathbb{Z}}$ are bases of two detail spaces $\{\mathbf{O}_j\}_{n\in\mathbb{Z}}$ and $\{\tilde{\mathbf{O}}_j\}_{n\in\mathbb{Z}}$ such that

$$\mathbf{V}_j \oplus \mathbf{O}_j = \mathbf{V}_{j-1}$$

and

$$\tilde{\mathbf{V}}_j \oplus \tilde{\mathbf{O}}_j = \tilde{\mathbf{V}}_{j-1}.$$

The biorthogonality of the decomposition and reconstruction wavelets implies that $\mathbf{O}_j$ is not orthogonal to $\mathbf{V}_j$, but is to $\tilde{\mathbf{V}}_j$, whereas $\tilde{\mathbf{O}}_j$ is not orthogonal to $\tilde{\mathbf{V}}_j$, but is to $\mathbf{V}_j$. The support size, the number of vanishing moments, the regularity and the symmetry of biorthogonal wavelets is controlled with an appropriate design of h and $\tilde{h}$.

Similar to the case of orthogonal wavelets, one can show that if $h[n]$ and $\tilde{h}[n]$ are nonzero respectively for $N_1 \leq n \leq N_2$ and $\tilde{N}_1 \leq n \leq \tilde{N}_2$, then ϕ and $\tilde{\phi}$ have a support equal to $[N_1, N_2]$ and $[\tilde{N}_1, \tilde{N}_2]$ respectively. By Equation (4.46), notice that

$$g[n] = (-1)^{1-n}\tilde{h}[1 - n], \ \ \text{and} \ \ \tilde{g}[n] = (-1)^{1-n}h[1 - n]$$

the support of ψ and $\tilde{\psi}$ are respectively $[(N_1 - \tilde{N}_2 + 1)/2, \ (N_2 - \tilde{N}_1 + 1)/2]$ and $[(\tilde{N}_1 - N_2 + 1)/2, \ (\tilde{N}_2 - N_1 + 1)/2]$. Both wavelets thus have the same size of support, which is equal to $(N_2 - N_1 + \tilde{N}_2 - \tilde{N}_1)/2$.

Since $\hat{\phi}(0) \neq 0$ and $\hat{\tilde{\phi}}(0) \neq 0$, Equation (4.25) and (4.49) show that the number of vanishing moments of ψ and $\tilde{\psi}$ depends on the number of zeros at $\omega = 0$ of $\hat{g}(\omega)$ and $\hat{\tilde{g}}(\omega)$ respectively. Notice $\hat{g}(\omega) = ae^{-i(2l+1)\omega}\hat{\tilde{h}}^*(\omega + \pi)$ and $\hat{\tilde{g}}(\omega) = a^{-1}e^{-i(2l+1)\omega}\hat{h}^*(\omega + \pi)$ for some $a \in \mathbb{R}$ and $l \in \mathbb{Z}$, we conclude that ψ has $\tilde{p}$ vanishing moments if and only if $\hat{\tilde{h}}(\omega)$ has a zero of order $\tilde{p}$ at $\omega = \pi$, whereas, $\tilde{\psi}$ has p vanishing moments if and only if $\hat{h}(\omega)$ has a zero of order p at $\omega = \pi$. On the other hand, the smoothness of ϕ and ψ can be related to the order of zeros of $\hat{h}(\omega)$ at $\omega = \pi$ [70]. This is intuitively make sense by Equation (4.48) and (4.49), the more number of zeros of $\hat{h}(\omega)$ at $\omega = \pi$, the smoother ϕ and ψ will be. The following remark summarizes the relationship between the number of vanishing moments and the regularity of biorthogonal wavelets.

Remark: *Let $\{h, g\}$ and $\{\tilde{h}, \tilde{g}\}$ be perfect reconstruction filter banks, and $\{\phi, \psi\}$ and $\{\tilde{\phi}, \tilde{\psi}\}$ are biorthogonal scaling functions and wavelets generated from $\{h, g\}$ and $\{\tilde{h}, \tilde{g}\}$. Let the order of zeros of $\hat{h}(\omega)$ and $\hat{\tilde{h}}(\omega)$ at $\omega = \pi$ be p and $\tilde{p}$ respectively, then the regularity of ϕ and ψ increases with p, which is the vanishing moments of $\tilde{\psi}$, similarly, the regularity of $\tilde{\phi}$ and $\tilde{\psi}$ increases with $\tilde{p}$, which is the vanishing moments of ψ.*

n	$p, \tilde{p}$	$h[n]$	$\tilde{h}[n]$
0		$\sqrt{2}/2$	$45\sqrt{2}/64$
$-1, 1$	$p = 2$	$\sqrt{2}/4$	$19\sqrt{2}/64$
$-2, 2$	$\tilde{p} = 2$		$-\sqrt{2}/8$
$-3, 3$			$-3\sqrt{2}/64$
$-4, 4$			$3\sqrt{2}/128$
$0, 1$		$3\sqrt{2}/8$	$11025\sqrt{2}/16384$
$-1, 2$	$p = 2$	$\sqrt{2}/8$	$-307\sqrt{2}/16384$
$-2, 3$	$\tilde{p} = 2$		$-3489\sqrt{2}/16384$
$-3, 4$			$363\sqrt{2}/16384$
$-4, 5$			$865\sqrt{2}/16384$
$-5, 6$			$-195\sqrt{2}/16384$
$-6, 7$			$-105\sqrt{2}/16384$
$-7, 8$			$35\sqrt{2}/16384$

Table 4.2: Perfect reconstruction filters h and $\tilde{h}$ for compactly supported spline biorthogonal wavelets with p and $\tilde{p}$ vanishing moments.

Since ψ and $\tilde{\psi}$ may not have the the same regularity and number of vanishing moments, the two reconstruction formulas

$$f = \sum_{j=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} < f, \psi_{j,n} > \tilde{\psi}_{j,n}$$

and

$$f = \sum_{j=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} < f, \tilde{\psi}_{j,n} > \psi_{j,n}$$

are not equivalent. To produce small wavelet coefficients in the regular regions, we must compute the inner products using the wavelet with the maximum number of vanishing moments. The reconstructions is then performed with the other wavelet, which is generally the smoother one.

It is possible to construct smooth biorthogonal wavelets of compact support which are

$\tilde{\phi}(t), (p = 2, \tilde{p} = 4)$ $\tilde{\psi}(t), (p = 2, \tilde{p} = 4)$ $\tilde{\phi}(t), (p = 3, \tilde{p} = 7)$ $\tilde{\psi}(t), (p = 3, \tilde{p} = 7)$

$\phi(t), (p = 2, \tilde{p} = 4)$ $\psi(t), (p = 2, \tilde{p} = 4)$ $\phi(t), (p = 3, \tilde{p} = 7)$ $\psi(t), (p = 3, \tilde{p} = 7)$
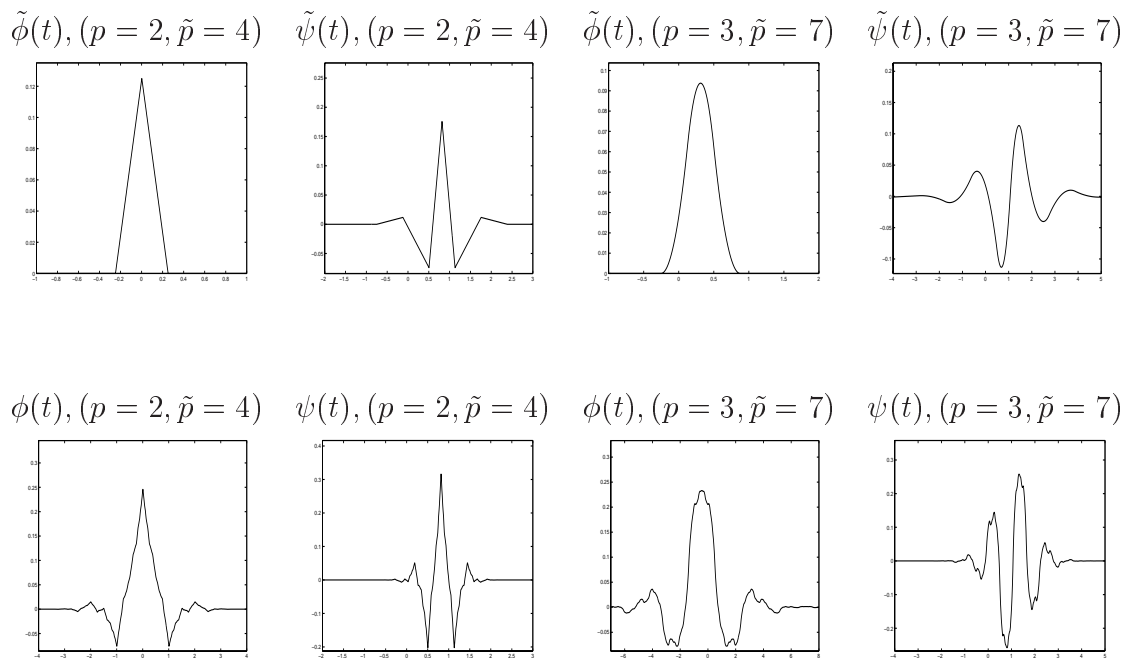
Figure 4.6: Spline biorthogonal scaling functions and wavelets for $(p = 2, \tilde{p} = 4)$ and $(p = 3, \tilde{p} = 7)$ at scale $j = 3$.

either symmetric or antisymmetric. This is impossible for orthogonal wavelets, except the particular case of the Haar basis. Symmetric or antisymmetric wavelets are synthesized with perfect reconstruction filters having a linear phase. This is a desirable property for many applications.

  The biorthogonal wavelets with a minimum size of support are constructed with a technique introduced in [78], which is similar to the construction of Daubechies wavelets. As an example, Table (4.2) gives the coefficients of perfect reconstruction filters of compactly supported spline wavelets for $(p = 2, \tilde{p} = 4)$ and $(p = 3, \tilde{p} = 7)$. The resulting symmetric and antisymmetric biorthogonal wavelets and scaling functions are shown in Figure (4.6).

4.4.3   Lifting Wavelets

A lifting is an elementary modification of perfect reconstruction filters, which is used to improve the wavelet properties. Compactly supported biorthogonal wavelet bases can be constructed from finite impulse response biorthogonal filters $\{h, g, \tilde{h}, \tilde{g}\}$ which satisfy

$$\hat{h}^*(\omega)\hat{\tilde{h}}(\omega) + \hat{h}^*(\omega + \pi)\hat{\tilde{h}}(\omega + \pi) = 2$$

and

$$\hat{g}(\omega) = e^{-i\omega}\hat{\tilde{h}}^*(\omega + \pi) \text{ and } \hat{\tilde{g}}(\omega) = e^{-i\omega}\hat{h}^*(\omega + \pi).$$

The filter h and $\tilde{h}$ are said to be dual. The following theorem [79] characterizes all filters of compact support that are dual to $\tilde{h}$.

Theorem 4.6 *Let h and $\tilde{h}$ be dual filters with a finite support. A filter h^l with finite support is dual to $\tilde{h}$ if and only if there exists a finite filter l such that*

$$\hat{h}^l(\omega) = \hat{h}(\omega) + e^{-i\omega}\hat{\tilde{h}}^*(\omega + \pi)\hat{l}^*(2\omega). \tag{4.52}$$

This theory proves that if $\{h, g, \tilde{h}, \tilde{g}\}$ are biorthogonal then we can construct a new set of biorthogonal filters $\{h^l, g, \tilde{h}, \tilde{g}^l\}$ with

$$\hat{h}^l(\omega) = \hat{h}(\omega) + \hat{g}(\omega)\hat{l}^*(2\omega) \tag{4.53}$$

and

$$\hat{\tilde{g}}^l(\omega) = e^{-i\omega}\hat{h}^{l*}(\omega + \pi) = \hat{\tilde{g}}(\omega) - \hat{\tilde{h}}(\omega)\hat{l}(2\omega). \tag{4.54}$$

The inverse Fourier transform of Equation (4.53) and (4.53) gives

$$h^l[n] = h[n] + \sum_{k=-\infty}^{+\infty} g[n - 2k]l[-k] \tag{4.55}$$

and

$$\tilde{g}^l[n] = \tilde{g}[n] - \sum_{k=-\infty}^{+\infty} \tilde{h}[n - 2k]l[k]. \tag{4.56}$$

The new filters are said to be lifted because the use of l can improve their properties. A new set of biorthogonal wavelet bases can be derived from the lifted filter banks [80].

(a)

(b)
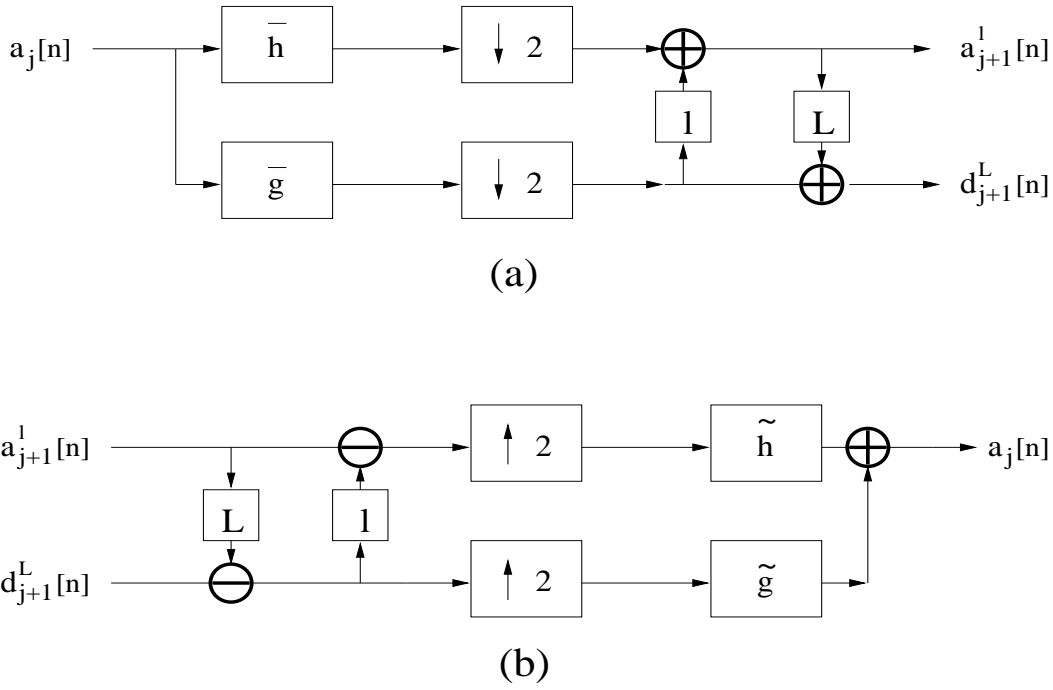
Figure 4.7: The biorthogonal filter banks with a lifting and a dual lifting.

Let $\{\phi, \psi, \tilde{\phi}, \tilde{\psi}\}$ be a family of compactly supported biorthogonal scaling functions and wavelets associated with the filter bank $\{h, g, \tilde{h}, \tilde{g}\}$. Let l be a finite sequence. A new family of biorthogonal scaling functions and wavelets $\{\phi^l, \psi^l, \tilde{\phi}, \tilde{\psi}^l\}$ is defined by

$$\phi^l(t) = \sqrt{2} \sum_{k=-\infty}^{+\infty} \left(h[k]\phi^l(2t - k) + l[-k]\psi^l(t - k) \right)$$

$$\psi^l(t) = \sqrt{2} \sum_{k=-\infty}^{+\infty} g[k]\phi^l(2t - k)$$

$$\tilde{\psi}^l(t) = \tilde{\psi}(t) - \sum_{k=-\infty}^{+\infty} l[k]\tilde{\phi}(2t - k). \tag{4.57}$$

If $\{\phi^l, \psi^l, \tilde{\phi}, \tilde{\psi}^l\}$ defined in Equation (4.57) have finite energy, then $\{\psi^l_{j,n}\}_{(j,n)\in\mathbb{Z}}$ and $\{\tilde{\psi}^l_{j,n}\}_{(j,n)\in\mathbb{Z}}$ are biorthogonal wavelet bases of $\mathbf{L}^2(\mathbb{Z})$.

The lifting increases the support size of ψ and $\tilde{\psi}$ typically by the length of the support of l. Design procedures compute minimum size filters l to achieve specific properties. Section (4.4.2) points out that the regularity of ϕ and ψ and the number of vanishing moments of $\tilde{\psi}$ is determined by the order of zeros of $\hat{\tilde{g}}(\omega)$ at $\omega = 0$. The coefficients of $l[n]$ are often calculated to produce a lifted transfer function $\hat{\tilde{g}}^l(\omega)$ with more zeros at

$\omega = 0$.

To increase the number of vanishing moments of ψ and the regularity of $\tilde{\phi}$ and $\tilde{\psi}$, we use a dual lifting which modifies $\tilde{h}$ and hence g instead of h and $\tilde{g}$. The corresponding lifting formula with a filter $L[k]$ are obtained by switching h with g and $\tilde{h}$ with $\tilde{g}$ in Equation (4.55) and (4.56)

$$g^L[n] = g[n] + \sum_{k=-\infty}^{+\infty} h[n - 2k]L[-k] \tag{4.58}$$

and

$$\tilde{h}^L[n] = \tilde{h}[n] - \sum_{k=-\infty}^{+\infty} \tilde{g}[n - 2k]L[k]. \tag{4.59}$$

The resulting family of biorthogonal scaling functions and wavelets $\{\phi, \psi^L, \tilde{\phi}^L, \tilde{\psi}^L\}$ can be constructed in the similar way in Equation (4.57).

Successive iteration of lifting and dual lifting can improve the regularity and vanishing moments of both ψ and $\tilde{\psi}$ by increasing the order of zeros of $\hat{g}(\omega)$ and $\hat{\tilde{g}}(\omega)$ at $\omega = 0$. A block diagram of biorthogonal filter banks with a lifting and a dual lifting is shown in Figure (4.7).

Example: Lazy Wavelets Lazy filters $\tilde{h}[n] = h[n] = \delta[n]$ and $\tilde{g}[n] = g[n] = \delta[n-1]$ satisfy the biorthogonality conditions (4.45). Their Fourier transforms are

$$\hat{\tilde{h}}(\omega) = \hat{h}(\omega) = 1 \ \ \text{and} \ \ \hat{\tilde{g}}(\omega) = \hat{g}(\omega) = e^{-i\omega}. \tag{4.60}$$

The resulting filter bank just separate the even and odd samples of a signal without filtering. The lazy scaling functions and wavelets associated with these filters are

$$\tilde{\phi}(t) = \phi(t) = \delta(t) \ \ \text{and} \ \ \tilde{\psi}(t) = \psi(t) = \delta(t - \frac{1}{2}).$$

Apparently they do not belong to $\mathbf{L}^2(\mathbb{R})$. These wavelets can be transformed into finite energy functions by appropriate liftings.

A lifting of a lazy filter $\hat{\tilde{g}}(\omega) = e^{-i\omega}$ yields

$$\hat{\tilde{g}}^l(\omega) = e^{-i\omega} - \hat{l}(2\omega).$$

To produce a symmetric wavelet $e^{i\omega}\hat{l}(2\omega)$ must be even. It can be verified [81] that the shortest l that lifts lazy wavelet to have 4 vanishing moments is defined by

$$\hat{l}(2\omega) = e^{-i\omega}\left(\frac{9}{8}\cos\omega - \frac{1}{8}\cos 3\omega\right).$$

Insert it in Equation (4.53) gives

$$\hat{h}^l(\omega) = -\frac{1}{16}e^{-3i\omega} + \frac{9}{16}e^{-i\omega} + 1 + \frac{9}{16}e^{i\omega} - \frac{1}{16}e^{3i\omega}.$$

The resulting ϕ^l is the Deslauriers-Dubuc [81] interpolating scaling function of order 4, and $\psi^l(t) = \sqrt{2}\phi^l(2t-1)$. Both ϕ^l and ψ^l are continuously differentiable but $\tilde{\phi}$ and $\tilde{\psi}^l$ are still sums of Diracs. A dual lifting can transform them into finite energy functions by creating a dual lifted filter $\hat{g}^L(\omega)$ with one or more zeros at $\omega = 0$.

Any biorthogonal filters $\{h, g, \tilde{h}, \tilde{g}\}$ can be synthesized with a succession of lifting and dual lifting applied to the lazy filters defined in Equation (4.60), up to shifting and multiplicative constants.

4.5 Separable Wavelet Bases

To any orthonormal wavelet basis $\{\psi_{j,n}\}_{(j,n)\in\mathbb{Z}^2}$ of $\mathbf{L}^2(\mathbb{R})$, one can associate a separable orthonormal basis of $\mathbf{L}^2(\mathbb{R}^2)$,

$$\{\psi_{j_1,n_1}(t_1)\psi_{j_2,n_2}(t_2)\}_{(j_1,n_1,j_2,n_2)\in\mathbb{Z}^4}.$$

The functions $\psi_{j_1,n_1}(t_1)\psi_{j_2,n_2}(t_2)$ mix information at two different scales 2^{j_1} and 2^{j_2} along t_1 and t_2, which we often want to avoid. Separable multiresolution leads to another construction of separable wavelet bases whose elements are products of one dimensional scaling functions and wavelets dilated at the same scale. These multiresolution approximations also have important applications in computer vision, where they are used to process images at different level of details. Lower resolution images are represented by fewer pixels and might still carry enough information to perform a recognition tasks.

Let $\{V_j\}, j \in \mathbb{Z}$ be a multiresolution approximation of $L^2(\mathbb{R})$. A separable two-dimensional multiresolution is composed of the tensor product spaces

$$V_j^2 = V_j \otimes V_j.$$

Theory (4.1) shows the existence of a scaling function ϕ such that $\{\phi_{j,n}\}_{n \in \mathbb{Z}}$ is an or-thonormal basis of V_j. By the classical theory of functional analysis, one can proves that for $t = (t_1, t_2)$ and $n = (n_1, n_2)$

$$\left\{\phi_{j,n}^2(t) = \phi_{j,n_1}(t_1)\phi_{j,n_2}(t_2)\right\}_{n \in \mathbb{Z}^2}$$

is an orthonormal basis of V_j^2. It is obtained by scaling the separable scaling function $\phi^2(t) = \phi(t_1)\phi_(t_2)$ and translating it onto a two dimensional grid with intervals of 2^j. Let W_j^2 be the detail space equal to the orthogonal complement of the lower resolution approximation space V_j^2 in V_{j-1}^2, i.e.,

$$V_{j-1}^2 = V_j^2 \oplus W_j^2.$$

To construct a wavelet orthonormal basis of $L^2(\mathbb{R}^2)$, the following theory [70] builds a wavelet basis of each detail space W_j^2.

Theorem 4.7 *Let ϕ be a scaling function and ψ be the corresponding wavelet generating a wavelet orthonormal basis of $L^2(\mathbb{R})$. We define three wavelets:*

$$\psi^1(t) = \phi(t_1)\psi(t_2), \quad \psi^2(t) = \psi(t_1)\phi(t_2), \quad \psi^3(t) = \psi(t_1)\psi(t_2), \tag{4.61}$$

and denote for $1 \le k \le 3$,

$$\psi_{j,n}^k(t) = \frac{1}{2^j}\psi^k\left(\frac{t_1 - 2^j n_1}{2^j}, \frac{t_2 - 2^j n_2}{2^j}\right). \tag{4.62}$$

The wavelet family

$$\{\psi_{j,n}^1, \psi_{j,n}^2, \psi_{j,n}^3\}_{n \in \mathbb{Z}^2}$$

is an orthonormal basis of W_j^2 *and*

$$\{\psi_{j,n}^1, \psi_{j,n}^2, \psi_{j,n}^3\}_{(j,n) \in \mathbb{Z}^3}$$

is an orthonormal basis of $L^2(\mathbb{R}^2)$.

(a) An image of toolbox

(b) 2D wavelet decomposition of toolbox
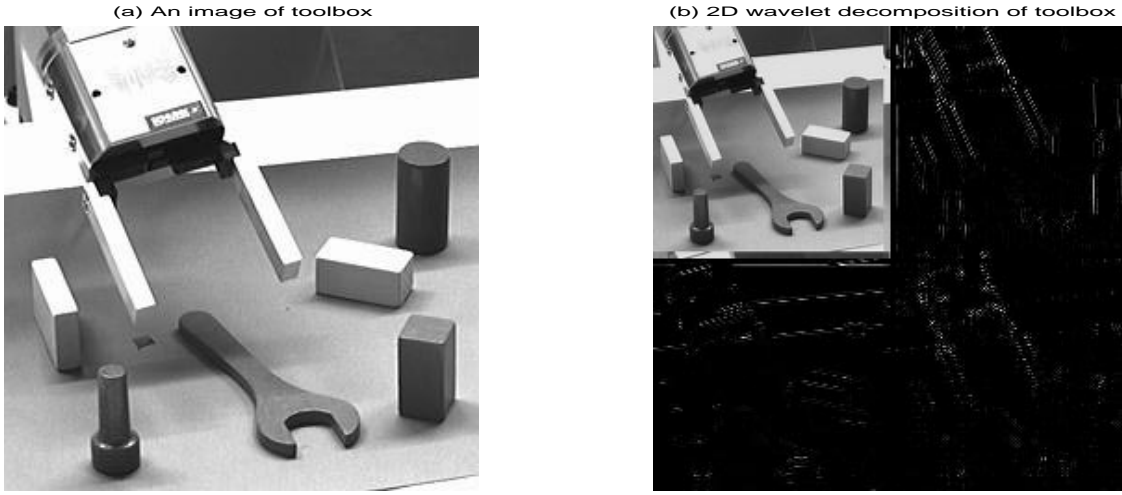
Figure 4.8: Wavelet decomposition of a toolbox image

The three wavelets extract image details at different scales and orientations. Over positive frequencies, $\hat{\phi}$ and $\hat{\psi}$ have an energy mainly concentrated respectively on lower and higher frequencies. Let $\boldsymbol{\omega} = (\omega_1, \omega_2)$, the separable wavelet expressions implies that

$$\hat{\psi}^1(\boldsymbol{\omega}) = \hat{\phi}(\omega_1)\hat{\psi}(\omega_2), \quad \hat{\psi}^2(\boldsymbol{\omega}) = \hat{\psi}(\omega_1)\hat{\phi}(\omega_2), \quad \hat{\psi}^3(\boldsymbol{\omega}) = \hat{\psi}(\omega_1)\hat{\psi}(\omega_2). \tag{4.63}$$

Hence $|\hat{\psi}^1(\boldsymbol{\omega})|$ is larger at low horizontal frequencies ω_1 and high vertical frequencies ω_2, $|\hat{\psi}^2(\boldsymbol{\omega})|$ is larger at high horizontal frequencies ω_1 and low vertical frequencies ω_2, whereas $|\hat{\psi}^3(\boldsymbol{\omega})|$ is larger at high horizontal frequencies ω_1 and high vertical frequencies ω_2. As a result, wavelet coefficients calculated with ψ^1 and ψ^2 are larger along edges which are respectively horizontal and vertical, and ψ^3 produces large coefficients at the corners. This is illustrated by the decomposition of a toolbox image in Figure (4.8).

In the similar fashion, one-dimensional biorthogonal wavelet bases can also be extended to separable biorthogonal bases of $\mathbf{L}^2(\mathbb{R}^2)$. let $\phi, \tilde{\phi}$ and $\psi, \tilde{\psi}$ be two dual pairs of scaling functions and wavelets that generate biorthogonal wavelet bases of $\mathbf{L}^2(\mathbb{R})$. The dual wavelets of ψ^1, ψ^2 and ψ^3 defined by Equation (4.61) are

$$\tilde{\psi}^1(\boldsymbol{t}) = \tilde{\phi}(t_1)\tilde{\psi}(t_2), \quad \tilde{\psi}^2(\boldsymbol{t}) = \tilde{\psi}(t_1)\tilde{\phi}(t_2), \quad \tilde{\psi}^3(\boldsymbol{t}) = \tilde{\psi}(t_1)\tilde{\psi}(t_2). \tag{4.64}$$

It is easy to verify that

$$\{\psi^1_{j,\boldsymbol{n}}, \psi^2_{j,\boldsymbol{n}}, \psi^3_{j,\boldsymbol{n}}\}_{(j,\boldsymbol{n})\in\mathbb{Z}^3}$$

and

$$\{\tilde{\psi}^1_{j,\boldsymbol{n}}, \tilde{\psi}^2_{j,\boldsymbol{n}}, \tilde{\psi}^3_{j,\boldsymbol{n}}\}_{(j,\boldsymbol{n})\in\mathbb{Z}^3}$$

are biorthogonal bases of $\mathbf{L}^2(\mathbb{R}^2)$.

It is possible to extend the fast one-dimensional wavelet transform algorithm to two dimensions. Let $f \in \mathbf{L}^2(\mathbb{R}^2)$, at all the scales of 2^j and for any $\boldsymbol{n} = (n_1, n_2)$, we denote

$$a_j[\boldsymbol{n}] =< f, \phi^2_{j,\boldsymbol{n}} >$$

and

$$d^k_j[\boldsymbol{n}] =< f, \psi^k_{j,\boldsymbol{n}} >, \ \ k = 1, 2, 3.$$

For any pair of one-dimensional filters $y[n]$ and $z[n]$, we write the product filter $yz[\boldsymbol{n}] = y[n_1]z[n_2]$, and denote $\bar{y}[n] = y[-n]$.

The wavelet coefficients at the scale 2^{j+1} are calculated from a_j with two dimensional separable convolutions and subsamplings. Let h and g be the conjugate mirror filters associated to the wavelet ψ. The decomposition formula are obtained by applying the one-dimensional convolutional formula to the separable two-dimensional wavelets and scaling functions. For $\boldsymbol{n} = (n_1, n_2)$,

$$a_{j+1}[\boldsymbol{n}] = a_j \star \bar{h}\bar{h}[\boldsymbol{n}], \tag{4.65}$$

$$d^1_{j+1}[\boldsymbol{n}] = a_j \star \bar{h}\bar{g}[\boldsymbol{n}], \tag{4.66}$$

$$d^2_{j+1}[\boldsymbol{n}] = a_j \star \bar{g}\bar{h}[\boldsymbol{n}], \tag{4.67}$$

$$d^3_{j+1}[\boldsymbol{n}] = a_j \star \bar{g}\bar{g}[\boldsymbol{n}]. \tag{4.68}$$

A separable two dimensional convolution can be factored into one-dimensional convolutions along with rows and columns of the images. The factorization is illustrated in Figure (4.9a). The rows of a_j are first convolved with $\bar{h}$ and $\bar{g}$, and subsampled by 2. The columns of these two output images are then convolved respectively with $\bar{h}$ and $\bar{g}$ and subsampled, which gives four subsampled images a_{j+1}, d^1_{j+1}, d^2_{j+1} and d^3_{j+1}.

We denote $\breve{y}[\boldsymbol{n}] = \breve{y}[n_1, n_2]$ the image obtained by inserting a row of zeros and a column of zeros between pairs of consecutive rows and columns of $y[n1, n2]$. a_j is recovered from the coarser scale approximation a_{j+1} and the wavelet coefficients d^1_{j+1}, d^2_{j+1} and

Rows Columns

(a)

Columns Rows

(b)
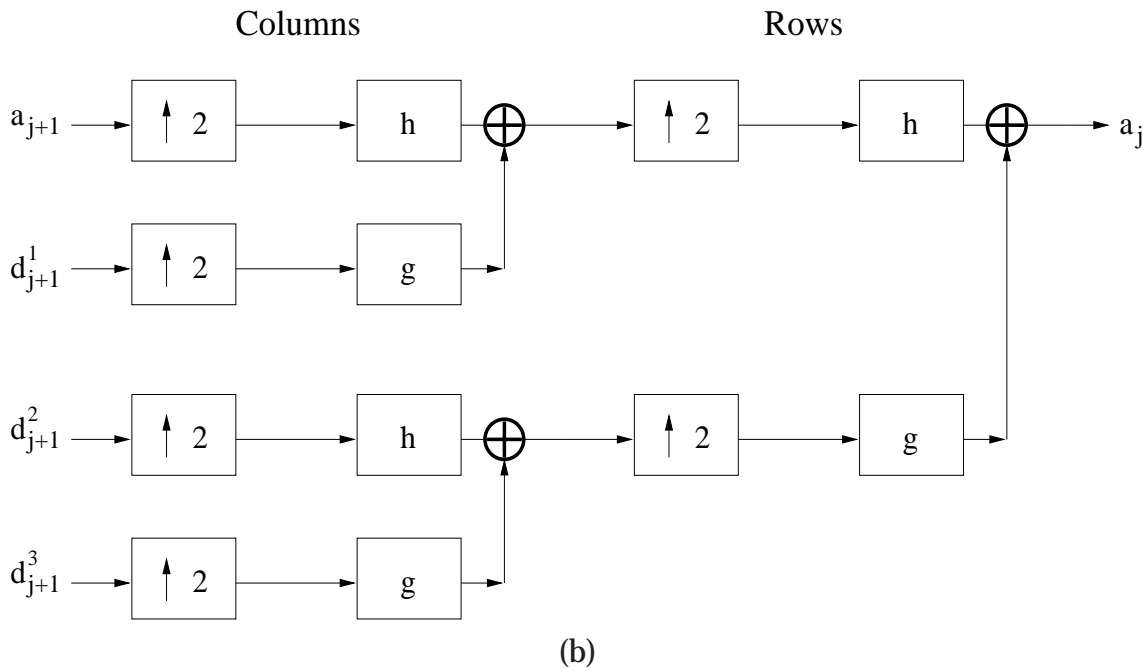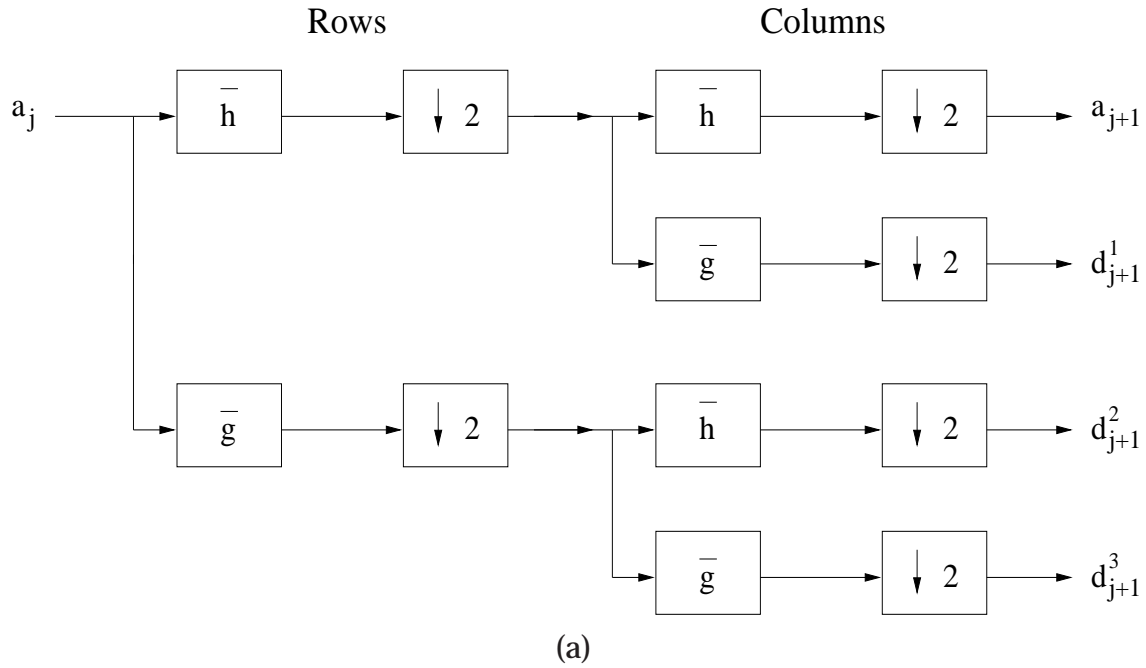
Figure 4.9: (a) A two-dimensional fast wavelet transform is computed with a cascade of filters $\bar{h}$ and $\bar{g}$ followed by a factor 2 subsampling in rows and columns respectively. (b) A two-dimensional fast inverse wavelet transform reconstructs progressively each a_j by inserting zeros between samples of a_{j+1} and d^k_{j+1}, $k = 1, 2, 3$, filtering and adding the outputs along with rows and columns.

d_{j+1}^3 with two-dimensional separable convolutions derived from the one-dimensional reconstruction formula in Equation (4.37)

$$a_j[\boldsymbol{n}] = \breve{a}_{j+1} \star hh[\boldsymbol{n}] + \breve{d}_{j+1}^1 \star hg[\boldsymbol{n}] + \breve{d}_{j+1}^2 \star gh[\boldsymbol{n}] + \breve{d}_{j+1}^3 \star gg[\boldsymbol{n}]. \tag{4.69}$$

These four convolutions can also be factored into six groups of one-dimensional convolutions along rows and columns, as illustrated in Figure (4.9b).

Let $b[\boldsymbol{n}]$ be an image whose pixels have a distance $2^J = N^{-1}$. We associate to $b[\boldsymbol{n}]$ a function $f(\boldsymbol{x}) \in V_J^2$ approximated at the scale 2^J. Its coefficients $a_J[\boldsymbol{n}] = < f, \phi_{J,\boldsymbol{n}}^2 >$ are discrete samples at scale of 2^J

$$b[\boldsymbol{n}] = N a_J[\boldsymbol{n}] \approx f(N^{-1}n).$$

The wavelet image representation of a_J is computed by iterating Equation (4.65-4.68) for $J < j \le L$:

$$\left[\, \{d_j^1, d_j^2, d_j^3\}_{J < j \le L}, a_L \, \right]$$

The original image a_J is recovered from this wavelet representation by iterating the reconstruction Equation (4.69) for $J < j \le L$.

4.6   Signal Estimation in Wavelet Framework

In this section, we consider the problem of signal estimation in an additive noise model. A signal $f \in I^2(\mathbb{Z})$ of support size N is contaminated by the addition of a noise. This noise is modeled by the realization of a zero mean random process W. The measured data are

$$Z[n] = f[n] + W[n], \;\; n \in \mathbb{Z}. \tag{4.70}$$

The signal f is estimated by transforming the noisy observation Z with a decision operator D, which is given by

$$\hat{f} = DZ. \tag{4.71}$$

A statistical approach usually assumes the knowledge of at least the probability distribution of the noise process W. An optimal D is to then minimize the risk of the estimator, which is the average loss calculated with respect to the probability distribution of noise

$$r(D, f) = E\{\|f - DZ\|\}. \tag{4.72}$$

Linear operators have long predominated the solution to this problem because of their simplicity, despite their limited performance.

The Bayes framework supposes that signals f are realizations of a random vector F whose probability distribution π is a known prior. The Bayes risk is the expected risk calculated with respect to the prior probability distribution π of the signal:

$$r(D, \pi) = E_\pi \left(r(D, f) \right). \tag{4.73}$$

Then the Bayes estimation is to optimize D to minimize the expected risk $r(D, \pi)$. It is, however, generally not possible to have enough information to define this prior probability distribution for a signal set with a complex structure. To overcome this difficulty, one may call upon a minimax framework that applies a simpler model which constrains signals in a prior set Θ. The goal is to then find an optimal operator which minimizes the maximum risk over Θ, i.e.,

$$D = \arg \inf_D r(D, \Theta), \tag{4.74}$$

where the maximum risk is given by

$$r(D, \Theta) = \sup_{f \in \Theta} r(D, f). \tag{4.75}$$

Except for a few special cases, minimax optimal operators are highly nonlinear and difficult to find for real world applications. More often than not, one settles for a suboptimal estimator. This section studies particular estimators that are diagonal in an orthonormal basis $B = \{g_m\}_{0 \leq m < N}$. If the basis B defines a sparse signal representation, then such diagonal estimators are nearly optimal among all nonlinear estimators. The noisy data

$$X = f + W \tag{4.76}$$

is decomposed in B. We write

$$X_B[m] =< X, g_m >, \ \ f_B[m] =< f, g_m >, \ \ W_B[m] =< W, g_m > .$$

The inner product of Equation (4.76) with g_m gives

$$X_B[m] = f_B[m] + W_B[m], \ 0 \le m < N.$$

We assume that W is a zero-mean white noise of variance σ^2, then

$$
\begin{aligned}
E\{W_B[m]W_B[p]\} &= \sum_{n=0}^{N-1}\sum_{k=0}^{N-1} g_m[n]g_p[k]E\{W[n]W[k]\} \\
&= \sigma^2 < g_p, g_m > \\
&= \sigma^2 \delta(p - m).
\end{aligned}
$$

The noise coefficient is hence also a white noise of variance σ^2.

A diagonal operator estimates independently each $f_B[m]$ from $X_B[m]$ with a function $d_m(x)$. The resulting estimator is

$$\hat{f} = DX = \sum_{m=0}^{N-1} d_m(X_B[m])g_m. \tag{4.77}$$

If setting $d_m(0) = 0$, we can write

$$d_m(X_B[m]) = a[m]X_B[m], \ \ 0 \le m < N$$

where $a[m]$ depends on $X_B[m]$. The estimation risk

$$r(D, f) = \sum_{m=0}^{N-1} E\{|f_B[m] - a[m]X_B[m]|^2\} = \sum_{m=0}^{N-1} |f_B[m]|^2(1 - a[m])^2 + \sigma^2 a[m]^2, \tag{4.78}$$

is minimized by

$$a[m] = \frac{|f_B[m]|^2}{|f_B[m]|^2 + \sigma^2} \tag{4.79}$$

and the minimum risk is

$$r_{\inf}(f) = \sum_{m=0}^{N-1} \frac{|f_B[m]|^2\sigma^2}{|f_B[m]|^2 + \sigma^2}. \tag{4.80}$$

In practice, the attenuation factor $a[m]$ in Equation (4.79) can not be computed since it depends on $|f_B[m]|$, whose value is not known. We hence refers to Equation (4.79) as

an oracle attenuation.

The analysis of diagonal estimators can be simplified by restricting $a[m] \in \{0,1\}$. A non-linear projector that minimizes the risk in Equation (4.78) is defined by

$$a[m] = \begin{cases} 1, & |f_B[m]| \geq \sigma \\ 0, & \text{otherwise} \end{cases}.$$

Similar to the case of Equation (4.79), this projector can not be implemented because $a[m]$ depends on the value of $|f_B[m]|$. The risk of this oracle projector is computed with Equation (4.78), and we obtain,

$$r_p(f) = \sum_{m=0}^{N-1} \min(|f_B[m]|^2, \sigma^2). \tag{4.81}$$

Since for any $x, y > 0$,

$$\min(x, y) \geq \frac{xy}{x+y} \geq \frac{1}{2}\min(x, y),$$

we have

$$r_p(f) \geq r_{\text{inf}}(f) \geq \frac{1}{2}r_p(f).$$

The risk of the oracle projector r_p is of the same order as the risk of the oracle attenuation (4.80). One can use the risk in (4.81) to verify the performance of practical thresholding estimators.

Instead of depending on $|f_B[m]|$, a feasible approach is to use $|X_B[m]|$ to determine an appropriate projection. A diagonal estimator can be written as

$$\hat{f} = DX = \sum_{m=0}^{N-1} d_m(X_B[m])g_m.$$

A hard thresholding estimator is implemented with

$$d_m(x) = \rho_T(x) = \begin{cases} x, & |x| \geq T \\ 0, & |x| < T \end{cases}.$$

The risk of this thresholding is

$$r_t(f) = \sum_{m=0}^{N-1} E\{|f_B[m] - \rho_T(X_B[m])|^2\}. \tag{4.82}$$

Since $X_B[m] = f_B[m] + W_B[m]$,

$$E\{|f_B[m] - \rho_T(X_B[m])|^2\} = \begin{cases} \sigma^2, & |X_B[m]| \geq T \\ |f_B[m]|^2, & \text{otherwise} \end{cases}.$$

Hence the risk of the hard thresholding estimator is larger than the risk of the oracle projector,

$$r_t(f) \geq r_p(f) = \sum_{m=0}^{N-1} \min(|f_B[m]|^2, \sigma^2).$$

An oracle attenuation yields a risk r_{inf} that is smaller than the risk of $r_p(f)$ of an oracle projection, by slightly decreasing the amplitude for all coefficients in order to reduce the added noise. A similar attenuation, although non-optimal, is implemented by a soft thresholding, which decreases by T the amplitude of all noisy coefficients. This soft thresholding function is given by

$$d_m(x) = \rho_T(x) = \begin{cases} x - T, & x \geq T \\ x + T, & x \leq -T \\ 0, & |x| < T \end{cases}.$$

It is the solution that minimizes a quadratic distance to the data, penalized by an l^1 norm. Given the data $x[m]$, the vector $y[m]$ which minimizes

$$\sum_{m=0}^{N-1} |y[m] - x[m]|^2 + 2T \sum_{m=0}^{N-1} |y[m]|$$

is $y[m] = \rho_T(x[m])$. The threshold is generally chosen so that it is just above almost all the noise coefficients. Since W_B is a vector of N independent Gaussian random variables of variance σ^2. By taking $T = \sigma \sqrt{2 \log_e^N}$, one can show that

$$\lim_{N \to +\infty} P(\max_m(W_B[m]) \leq T) = 1.$$

The following theorem [82] proves that the risk of a thresholding estimator is close to the risk of an oracle projector r_p defined in Equation (4.81).

Theorem 4.8 *Let $T = \sigma \sqrt{2 \log_e^N}$. The risk of a hard or soft thresholding estimator satisfies for all $N \geq 4$*

$$r_t(f) \leq (2 \log_e^N + 1)(\sigma^2 + r_p(f)).$$

A filter bank tree of depth $L - J > 0$ decomposes a discrete signal in a discrete wavelet basis defined in Section 4.3.

$$B = \left[\ \{\phi_L[k - 2^{j-J}n]\}_{n \in \mathbb{Z}}, \ \{\psi_j[k - 2^{j-J}n]\}_{J < j \leq L, n \in \mathbb{Z}} \ \right].$$ (4.83)

is an orthonormal basis of $\mathbf{I}^2(\mathbb{Z})$. A wavelet thresholding estimator can be written as

$$\hat{f} = \sum_{j=J+1}^{L} \sum_{m=-\infty}^{+\infty} \rho_T(< X, \psi_{j,m} >)\psi_{j,m} + \sum_{m=-\infty}^{+\infty} \rho_T(< X, \phi_{L,m} >)\psi_{j,m}.$$ (4.84)

where $\rho_T(\cdot)$ is a hard or soft thresholding function. In a wavelet signal representation, large amplitude coefficients correspond to transient signal variations, this means that the thresholding estimation only keeps transients coming from the underlying signal, without adding others due to the noise.

The threshold $T = \sigma\sqrt{2\log_e^N}$ is not optimal and in general a lower threshold reduces the risk. A threshold adapted to the data is calculated by minimizing an estimation of the risk. Denote $r_t(f, T)$ the risk of a soft thresholding estimator calculated with a threshold T. An estimate $\hat{r}_t(f, T)$ of $r_t(f, T)$ is calculated from the noisy data X, and T is optimized by minimizing $\hat{r}_t(f, T)$. To estimate the risk $r_t(f, T)$, observe that if $|X_B[m]| < T$ then the soft thresholding sets this coefficient to zero, which produces a risk equal to $|f_B[m]|^2$. Since

$$E\{|X_B[m]|^2\} = |f_B[m]|^2 + \sigma^2,$$

one can estimate $|f_B[m]|^2$ with $|X_B[m]|^2 - \sigma^2$. If $|X_B[m]| \geq T$, the soft thresholding subtracts T from the amplitude of $X_B[m]$. The expected risk is the sum of the noise energy plus the bias introduced by the reduction of the amplitude of $X_B[m]$ by T. It is estimated by $\sigma^2 + T^2$. The resulting estimator of $r_t(f, T)$ is

$$\hat{r}_t(f, T) = \sum_{m=0}^{N-1} \Phi(|X_B[m]|^2)$$ (4.85)

with

$$\Phi(u) = \begin{cases} u - \sigma^2, & u \leq T^2 \\ \sigma^2 + T^2, & u > T^2 \end{cases}.$$

It can be shown [82] that $\hat{r}_t(f, T)$ is a Stein Unbiased Risk Estimator (SURE) of $r_t(f, T)$, i.e. $E\{\hat{r}_t(f, T)\} = r_t(f, T)$.

To find the $\hat{T}$ that minimizes the SURE estimator $\hat{r}_t(f, T)$, the N data coefficients $X_B[m]$ are sorted in decreasing amplitude order with $O(N \log_2 N)$ operations. Let $X_B^r[k] = X_B[m_k]$ be the coefficient of rank $k : |X_B^r[k] \geq |X_B^r[k+1]|$ for $0 \leq k < N - 2$. Let l be the index such that $|X_B^r[l+1]| \leq T < |X_B^r[l]|$. We can rewrite Equation (4.85):

$$\hat{r}_t(f, T) = \sum_{k=l}^{N-1} |X_B^r[k]|^2 - (N - l)\sigma^2 + l(\sigma^2 + T^2). \qquad (4.86)$$

To minimize $\hat{r}_t(f, T)$, we must choose $T = |X_B^r[l+1]|$ because $\hat{r}_t(f, T)$ is increasing in T. It is therefore sufficient to compare the N possible values, $\{|X_B^r[k]|\}_{0 \leq k < N}$, to find the $\hat{T}$ that minimizes $\hat{r}_t(f, T)$, that requires $O(N)$ operations if we progressively recompute the formula (4.86). The calculation of $\hat{T}$ is thus performed with $O(N \log_2 N)$ operations. Although the estimator $\hat{r}_t(f, T)$ of $r_t(f, T)$ is unbiased, its variance may induce errors due to the noise energy, especially when $\|f\|^2 \ll E\{\|W\|^2\} = N\sigma^2$. In this case, one must impose $T = \sigma\sqrt{2\log_e N}$ in order to remove all the noise. Since $E\{\|X\|^2\} = \|f\|^2 + N\sigma^2$, we estimate $\|f\|^2$ with $\|X\|^2 - N\sigma^2$ and compare this value with a minimum energy level $\epsilon_N = \sigma^2 N^{1/2}(\log_e N)^{3/2}$. The resulting SURE threshold is

$$T = \begin{cases} \sigma\sqrt{2\log_e N}, & \|X\|^2 - N\sigma^2 \leq \epsilon_N \\ \hat{T}, & \|X\|^2 - N\sigma^2 > \epsilon_N \end{cases}.$$

Let Θ be a signal set and $\min_T r_t(\Theta)$ be the minimax risk of a soft thresholding obtained by optimizing the choice of T depending on Θ. Donoho and Johnstone [82] prove that the threshold computed empirically with the above equations yields a nearly minimax risk.

Figure (4.10) demonstrates the estimation result of a noisy piecewise smooth signal with a soft threshold estimator with the SURE threshold T.
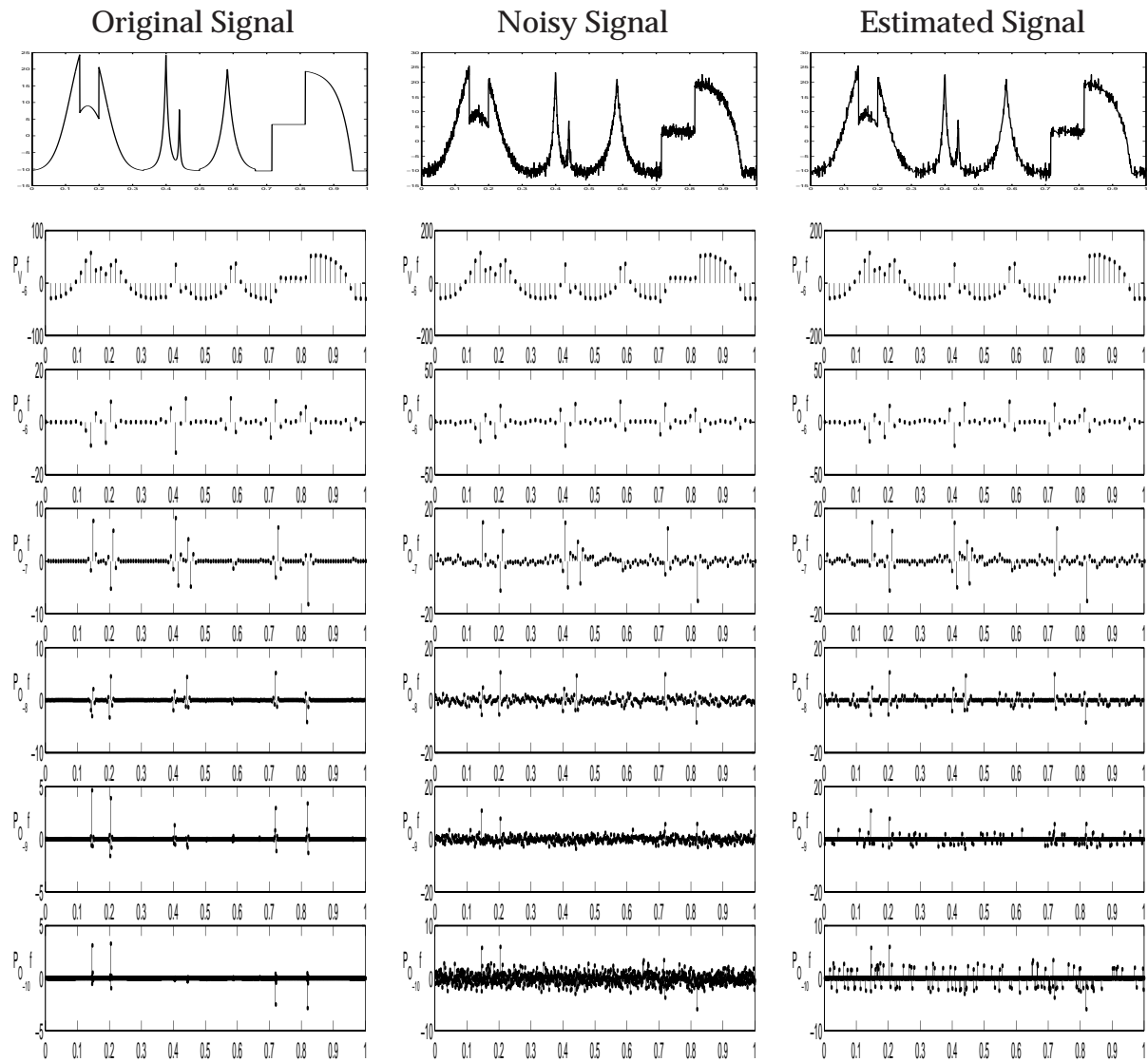
Figure 4.10: Estimation results of a noisy piecewise smooth signal with a soft threshold estimator with the SURE threshold T.

Multiscale Signal Enhancement: Beyond the Normality and Independence Assumption

C URRENT approaches to denoising or signal enhancement in wavelet-based framework have generally relied on the assumption of normally distributed perturbations. In practice, this assumption is often violated and sometimes, even prior information of probability distribution of the noise process is not available. To relax this assumption, we propose a novel non-linear filtering technique in this section. The key idea is to project a noisy signal onto a wavelet domain and to suppress wavelet coefficients by a mask derived from curvature extrema in its scale space representation. For a piecewise smooth signal, it can be shown that filtering by this curvature mask is equivalent to preserving the signal pointwise Hölder exponents at the singular points and lifting its smoothness at all the remaining points.

5.1   Introduction

Inspired by the fact that the human visual system processes and analyzes image information at different scales, researchers have made extensive use of the multiscale analysis in signal and image processing applications. Wavelet theory has played a particularly important role in multiscale analysis due to the fact that the basis functions are well suited to analyze local scale phenomena. This property also endows wavelets with a remarkable property for denoising in a wavelet based framework.

Donoho and Johnstone [7] first showed that effective noise suppression may be achieved by wavelet shrinkage. Given the noisy wavelet coefficients, i.e. the true wavelet coefficients plus a noise term, and assuming that one has knowledge of the true wavelet coefficients, an ideal filter sets a noisy coefficient to zero if the noise variance σ^2 is greater than the square of the true wavelet coefficient; otherwise the noisy coefficient is kept. In this way, the mean square error of this ideal estimator is the minimum of σ^2 and the square of the coefficient. Under the assumption of i.i.d. normal noise, it can shown that a soft thresholding estimator achieves a risk at most $O(\log M)$ times the risk of this ideal estimator, where M is the length of the observation.

To choose an appropriate threshold, Donoho and Johnstone [7] have taken a minimax approach to characterize the signal, and they proved, by setting a threshold $T = \sigma\sqrt{2\log_e M}$, that the estimation risk is close to the minimax risk. Krim and Pesquet [8] have given an alternative derivation for this threshold, using Rissanen's Minimum Description Length (MDL) criterion [9] and the assumption of normally distributed noise. The threshold T increase with M is due to the tail of the Gaussian distribution, which tends to generate larger noise coefficients when sample size increases. This threshold is not optimal, and in general a lower threshold reduces the risk. To refine the threshold, a SureShrink [10] procedure is proposed. Sureshrink calculates thresholds by the principle of minimizing the Stein unbiased estimate of risk for threshold estimates. SureShrink is also based on the assumption of i.i.d. normal noise. For non-Gaussian type of noise, Neumann [11], Averkamp and Houdre [12] studied the choice of thresholds by having recourse to asymptotics. Wavelet thresholding theory is, how-

ever, based on the assumption that we know the statistics of the noise to determine an adequate threshold. This makes the algorithm less flexible and less adaptive to different scenarios which can result in an even worse reconstruction. Compensation for the lack of a prior knowledge of the noise statistics may be handled by adopting the minimax principle [13] upon deriving the worst case noise distribution.

Points of sharp variations are often among the most important features for analyzing properties of transient piecewise smooth signals. To characterize the singular structures, Hölder exponents [14] provide a pointwise measure of a function over a time interval. Due to the pioneering work by Jaffard [15] and Meyer [16], it can be shown that a local signal singularity of a function is characterized by the decay of its wavelet transform amplitude across scales.

In this paper we first reinterpret the denoising problem as one of having to adjust the pointwise smoothness of noisy data, and propose a novel non-linear estimator as a result. The key idea is to separate out the signal portion from its noisy data to preserve the original smoothness property of the underlying signal, while the remaining noisy data admits the same Hölder exponents as the noise. This non-linear filter would be optimal in the sense of recovering the smoothness of the true underlying signal.

A crucial step for realizing such a smoothness-constrained filter is to identify the singularities of the true signal. We tackle this problem with the theory of curve evolution [83, 84], which is inherently geometric in nature, and widely used in computer vision [85, 86] and image processing [87]. The basic idea is that a planar curve deforms in the direction of its Euclidean normal, with a speed equal to its curvature. The noise riding on the signal has relatively higher curvatures in comparison to the underlying signal. It thus tends to be smoothed much faster than the latter. This disparity in evolution speed [38] is key to preserving the true features of the signal.

With the knowledge of the singularities, we proceed to generate a multiscale curvature mask to filter the wavelet transform of the noisy data. Specifically, we prove that filtering the transform by a curvature mask is equivalent to keeping the pointwise Hölder exponent of the noisy data at singular points, and lifting its smoothness at all the re-

maining points. In addition, residual data (noisy data minus estimate) admits the same Hölder exponents as the noise except at the singular points of the signal.

In the section that follows, we briefly introduce the concept of Hölder exponent and regularity measurement with wavelets. In Section 5.3, we give a concise statement of the problem. In Section 5.4, we derive the singularity detection algorithm through curve evolution. In Section 5.5, we formally define the smoothness-constrained filter, propose its implementation and verify its properties. Some numerical results appear in Section 5.6. Finally, we provide concluding remarks in Section 5.7.

5.2 Regularity measurement with wavelets

To characterize singular structures, it is necessary to precisely quantify the local regularity of a signal. Hölder spaces and Hölder exponent provide a uniform regularity measurement over time intervals, as well as at a particular point.

Definition 5.1 *Let $I \subseteq \mathbb{R}$ and f be a continuous function from I to $\mathbb{R}$. f is said to belong to a global Hölder space $C^{\alpha}(I)$, $\alpha > 0$ if and only if for any $v \in I$ there exists a positive constant c and a polynomial P_v of degree $m = \lfloor \alpha \rfloor$, $\lfloor \alpha \rfloor$ denotes the largest integer $m \leq \alpha$, such that*

$$|f(x) - P_v(x - v)| \leq c|x - v|^{\alpha}, \ \ \forall x \in I.$$

Definition 5.2 *f is said to belong to a pointwise Hölder space $C^{\alpha}(x_0)$, $\alpha > 0$, $x_0 \in I$ if and only if there exists a positive constant c and a polynomial P_{x_0} of degree $m = \lfloor \alpha \rfloor$ such that*

$$|f(x) - P_{x_0}(x - x_0)| \leq c|x - x_0|^{\alpha}, \ \ \forall x \in I.$$

Definition 5.3 *A function f is said to have a Hölder exponent α at point x_0 if there exists a polynomial P_{x_0} of degree $m = \lfloor \alpha \rfloor$ and f satisfies*

- *for any $\beta < \alpha$,*
$$\lim_{\Delta \to 0} \frac{|f(x_0 + \Delta) - P_{x_0}(\Delta)|}{|\Delta|^{\beta}} = 0$$

- *if $\alpha < +\infty$, for any $\beta > \alpha$*

$$\lim_{\Delta \to 0} \frac{|f(x_0 + \Delta) - P_{x_0}(\Delta)|}{|\Delta|^\beta} = +\infty.$$

The vanishing moment property of a wavelet function is crucial to measure the local regularity of a signal. If a wavelet $\psi(\cdot)$ has n vanishing moments, i.e.

$$\int_{-\infty}^{+\infty} t^k \psi(t) dt = 0, \ \ 0 \le k < n, \tag{5.1}$$

it can be shown [70] that the wavelet transformation is actually a multiscale differential operator of order n. This nice property relates the differentiability of f with its wavelet transform decay at fine scales.

Due to the pioneering work by Meyer [16] and Jaffard [15], it can be shown that a local signal singularity is characterized by the decay of its wavelet transform amplitude across scales.

Theorem 5.1 *[15] Let $\psi(\cdot)$ be a wavelet with n vanishing moments, $f \in \mathbf{L}^2(\mathbb{R})$ and $Wf(\cdot, \cdot)$ denotes its wavelet transform. Suppose f has a Hölder exponent $\alpha < n$ at x, then there exists a constant A such that*

$$\forall (u, s) \in \mathbb{R} \times \mathbb{R}^+, |Wf(u, s)| \le A s^{\alpha + \frac{1}{2}}(1 + |\frac{u - x}{s}|^\alpha).$$

Conversely, if $\alpha < n$ is not an integer and there exists a constant A and $\alpha' < \alpha$ such that

$$\forall (u, s) \in \mathbb{R} \times \mathbb{R}^+, |Wf(u, s)| \le A s^{\alpha + \frac{1}{2}}(1 + |\frac{u - x}{s}|^{\alpha'}).$$

then f is Hölder α at x.

5.3 Problem Formulation

A signal $f \in \mathbf{L}^2(\mathbb{R})$ is contaminated by the addition of a noise. This noise is modeled by the realization of a zero mean random process $\{N\}$. The measured data are

$$Z(x) = f(x) + N(x), \ \ x \in \mathbb{R}. \tag{5.2}$$

The signal f is estimated by transforming the noisy observation Z with a decision operator D, which is given by

$$\hat{f} = DZ. \tag{5.3}$$

A statistical approach usually assumes the knowledge of at least the probability distribution of the noise process $\{N\}$, and maybe a prior distribution of the signal. An optimal D is to then minimize the risk of the estimator, which is the average loss calculated with respect to the probability distribution of noise

$$r(D, f) = E\{\phi(f - DZ)\}, \tag{5.4}$$

where $\phi(\cdot)$ is a cost function. It is, however, generally not possible to have enough information to define this prior probability distribution for a signal set with complex structure. To overcome this difficulty, one may call upon a minimax framework that applies a simpler model which constrains signals in a prior set Θ. The goal is to then find an optimal operator which minimizes the maximum risk over Θ, i.e.,

$$D = \arg\inf_{D} r(D, \Theta), \tag{5.5}$$

where the maximum risk is given by

$$r(D, \Theta) = \sup_{f \in \Theta} r(D, f). \tag{5.6}$$

Except for a few special cases, minimax optimal operators are highly nonlinear and difficult to find for real world applications. More often than not, one settles for a suboptimal estimator. The well known thresholding estimator in an orthonormal wavelet basis proposed by Donoho and Johnstone [10] has a suboptimal risk $r_t(\Theta) \sim (\log_e M) r_{\min}(\Theta)$ for the set of piecewise smooth signals, where M is the observation length and $r_{min}(\Theta)$ is the minimax risk.

In this paper, we propose a novel nonlinear filtering technique. Contrary to statistical methods, we assume that a prior knowledge about pointwise smoothness measure of the signal is known or can be extracted. However, this smoothness property of the signal is corrupted by additive noise, which in general has a uniform Hölder exponent less

than 1. As noted earlier, we view the denoising problem as one of carefully controlling the Hölder exponents of measured data with a goal of extracting the signal portion with some smoothness fidelity to the original. Let $\alpha_f(\cdot)$ and $\alpha_N(\cdot)$ characterize the pointwise Hölder exponent of $f(\cdot)$ and $N(\cdot)$ respectively. A ideal operator T satisfies the following two conditions:

- $\hat{f} = TZ$ admits $\alpha_f(\cdot)$ as its pointwise Hölder exponent

- $V(x) = Z(x) - \hat{f}(x)$ admits $\alpha_N(\cdot)$ as its pointwise Hölder exponent

This non-linear filter is optimal in the sense of recovering the smoothness of the true underlying signal.

5.4 Curve evolution and singularity detection

Singularities and irregular structures often carry the most important information in signals. Many researchers [88], [89], [90], [91] and [92] have developed singularity detection techniques based on multiscale transforms. In this Section, we discuss the evolution of a planar curve in Euclidean space, in which the planar curve evolves in the direction of its Euclidean normal, with a speed equal to its local curvature. Using the well known foundation of such an evolution, we proceed to derive the partial differential equations that characterize the evolution of curvature. We subsequently propose a new singularity detection method by tracking the curvature extrema across scales.

Let $\mathbf{C}(\cdot, 0) : \mathbf{S}^1 \to \mathbb{R}^2$ be a smooth planar curve in a Euclidean space, then a geometric curve flow [83] [93], $\mathbf{C}(s, t) : \mathbf{S}^1 \times [0, T) \to \mathbb{R}^2$, is characterized by the following partial differential equation

$$\frac{\partial \mathbf{C}}{\partial t} = k\mathcal{N}$$

where t denotes the scale, k is the curvature of C, and $\mathcal{N}$ is its unit normal vector. This geometric curve flow is illustrated in Figure (5.1).

If we restrict the curve in Cartesian coordinates in $\mathbb{R}^2$ so that $\mathbf{C} = (x, y)$ is locally
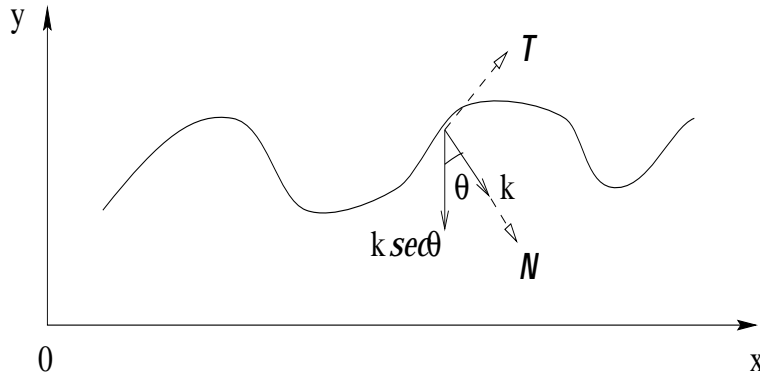
Figure 5.1: A geometric curve flow.

a graph, and require points to fix their x-coordinate during evolution, then we get a different flow $y(x,t) : \mathbb{R} \times [0,T) \to \mathbb{R}$ that can be represented as

$$\begin{cases} \frac{\partial y(x,t)}{\partial t} = k(x,t) \sec \theta(x,t) \\ y(x,0) = y_0(x) \end{cases} \tag{5.7}$$

where $y_0(\cdot)$ is an initial smooth curve in cartesian coordinates, $\theta(x,t) = \tan^{-1}(y'(x,t))$, " $'$ " denotes differentiation with respect to x, and $k(x,t)$ is the curvature given by

$$k = \frac{y''}{(1+y'^2)^{3/2}}. \tag{5.8}$$

The modification term for vertical speed is $\sec \theta$

$$\sec \theta = [(1+y'^2]^{1/2}. \tag{5.9}$$

Substituting Equations (5.8) and (5.9) into (5.7), we obtain an evolution of y for a fixed x,

$$\frac{\partial y}{\partial t} = \frac{y''}{1+y'^2}. \tag{5.10}$$

In addition, if $|y'|$ is bounded, Grayson [83] showed $k(x,t)$, as a result, also evolves according to equation

$$\frac{\partial k}{\partial t} = \frac{k''}{1+y'^2} + k^3. \tag{5.11}$$

Curvature is a natural indication of sharp variations in a signal. When a signal is contaminated by an additive noise, it makes sense to detect the singularities by tracking

the curvature extrema across curve evolution. Equation (5.11) indicates that the noise riding on the signal has relatively higher smoothing evolution speed in comparison to the underlying features. It thus tends to be "washed" away much faster than the latter. This disparity in evolution speed is key to extracting the true features of the signal. To obtain an accurate estimate of singularity, we also have to ensure that no new local extrema can be created during the curve evolution, which is generally referred to as the *Causality* property. The following proposition validates this requirement.

Proposition 5.1 *New extrema of y and k can not be created in passing from fine to coarse scales, i.e., for any $t_1 > t_0 > 0$, all the extrema points of $y(\cdot, t_1)$ and $k(\cdot, t_1)$ are extrema points in $y(\cdot, t_0)$ and $k(\cdot, t_0)$.*

Proof: Applying the Maximum Principle [94], if $|y'|$ is bounded, it follows directly that new maxima and minima of y and k can not be created as a curve evolves by equation (5.10). In fact, by lemma 1.9 in [83], we can show that for a given choice of cartesian coordinates, local minima of y and k increase with time, and local maxima decrease. Furthermore, the points of the curve where y and k reach their local minima and maxima vary continuously with time. ∎

A test signal defined on interval $[0, 1]$ is shown in Figure 5.2(a). By evolving it, we have the curvature variation across scales as illustrated in Figure 5.2(b), and the curvature extrema propagation line across evolution scales in Figure 5.2(c). Then we define a function $L(x)$ to denote the length of extrema propagation line originated at $x \in [0, 1]$ and a set $\Omega = \{x : L(x) > T\}$. T is a threshold and we set it to be half of the length of the longest propagation line of curvature extrema. Finally we define a singularity indicator function $p(\cdot)$ as

$$p(u) = \sum_{x_i \in \Omega} \delta(u - x_i). \qquad (5.12)$$
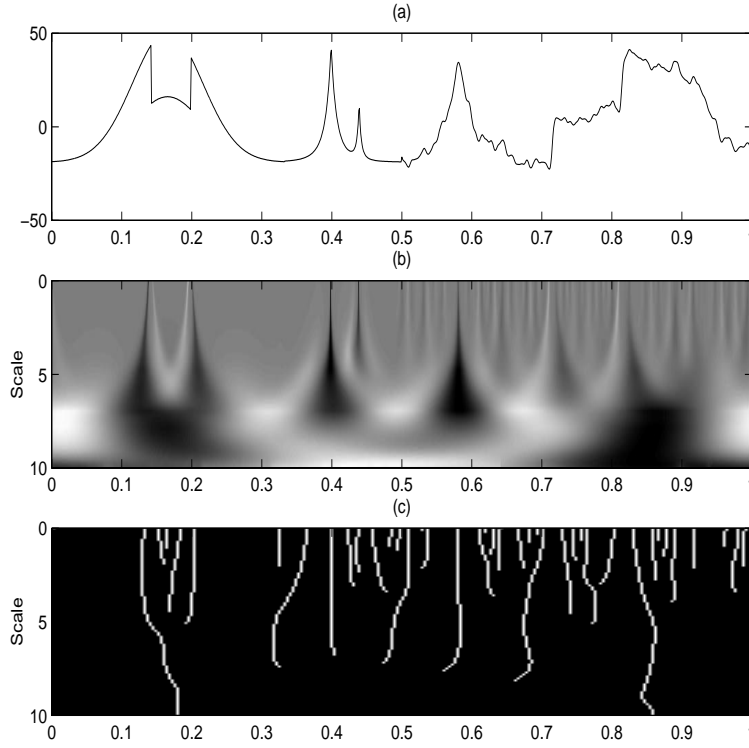
Figure 5.2: Singularity detection by tracking curvature extrema through curve evolution

5.5 A smoothness constrained filter

Let $f \in \mathbf{L}^2(\mathbb{R})$ be a piecewise smooth signal. Its observation $Z(\cdot)$ is modeled as in Equation (5.2). Let $\psi(x)$ be a compact support wavelet with n vanishing moments and $WZ(s, u)$ denote the continuous wavelet transformation of $Z(\cdot)$, which is given by

$$WZ(s, u) = Z(x) \star \psi_{s,u}(x) \tag{5.13}$$

where $\psi_{s,u}(x) = \frac{1}{\sqrt{s}}\psi\left(\frac{x-u}{s}\right)$.

By tracking curvature extrema through curve evolution of measured data $Z(\cdot)$, we get a singularity indicator function $p(\cdot)$ as defined in Equation (5.12).

Let

$$\rho(u) = \begin{cases} \exp\left(\frac{-u^2}{1-u^2}\right), & |u| < 1 \\ 0, & |u| \geq 1 \end{cases}. \tag{5.14}$$

and C be the support of $\psi(\cdot)$. Then we define
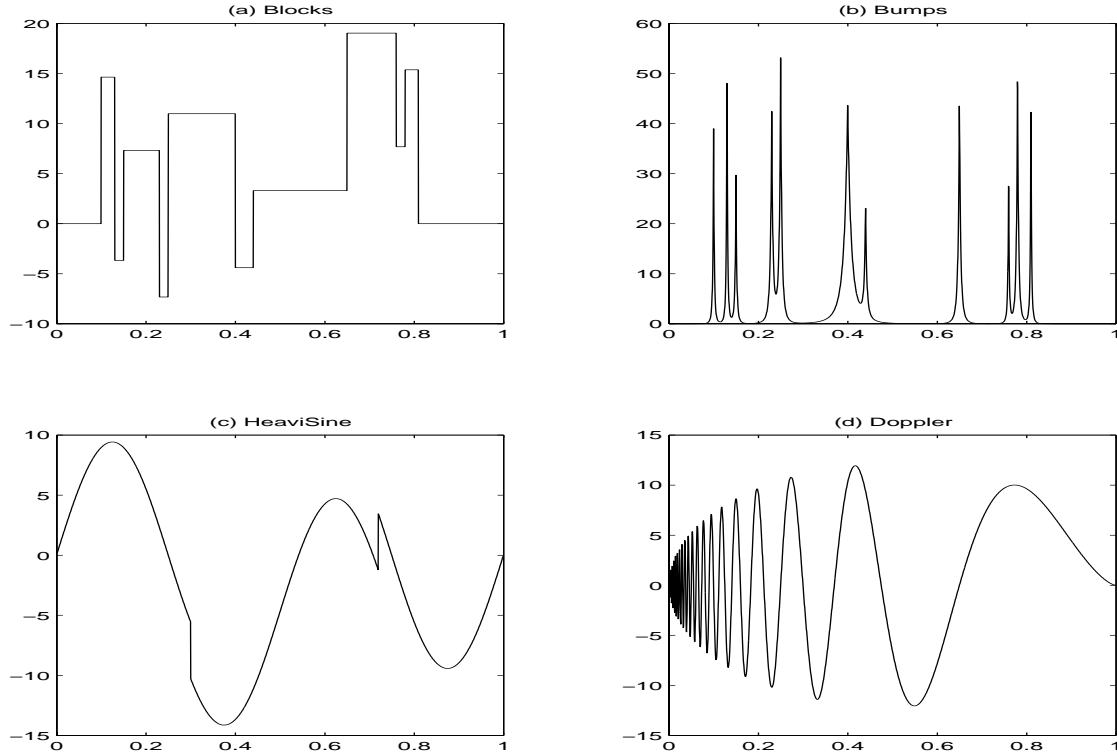
Figure 5.3: Clean signals of *Blocks, Bumps, HeaviSine and Doppler*

$$q(s, u) = p(u) \star \rho(\frac{u}{Cs}), \quad s > 0.$$

and a multiscale mask is given by

$$h(s, u) = \begin{cases} q(s, u), & \text{if } q(s, u) < 1 \\ 1, & \text{otherwise}. \end{cases} \tag{5.15}$$

Since this mask is derived from curvature extrema, we call it a multiscale curvature mask. Then a smoothness constrained filter T for piecewise smooth signals may be defined as

$$\hat{f}(x) = TZ(x) = \mathbf{W}^{-1}\{WZ(s, u)h(s, u)\}, \tag{5.16}$$

where $\mathbf{W}^{-1}$ denotes the inverse wavelet transform.

To analyze the property of this filter, we first investigate the pointwise Hölder exponents of the measured data $Z(\cdot)$.
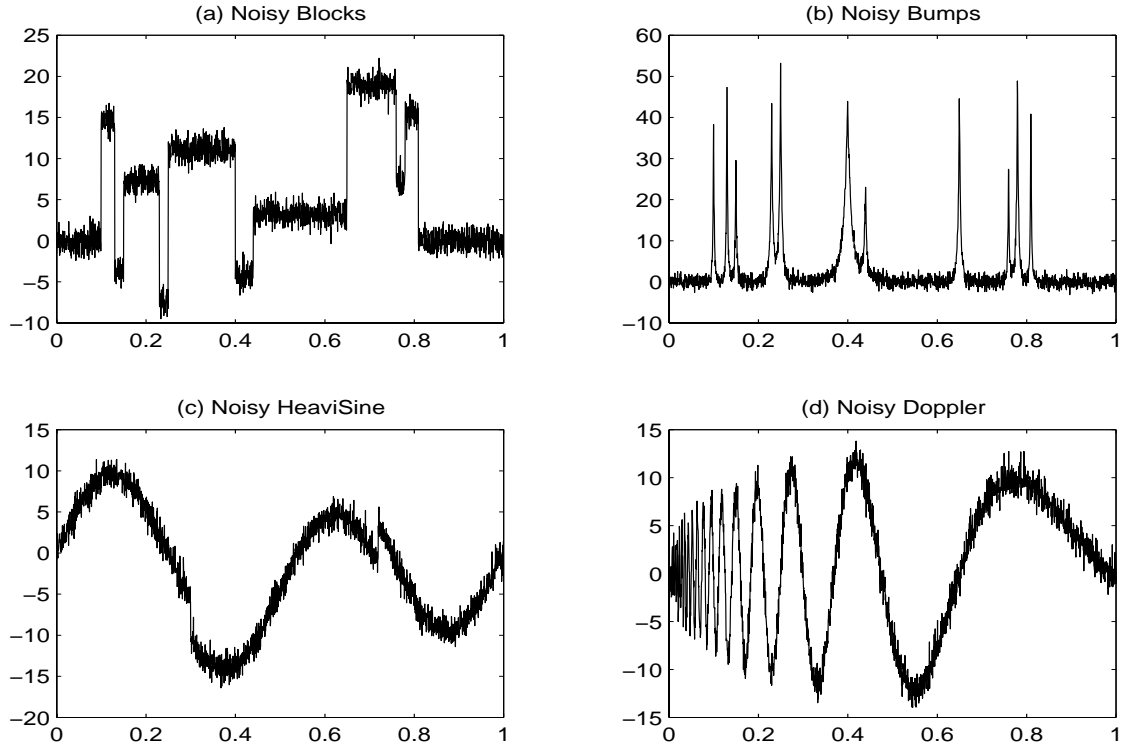
Figure 5.4: Gaussian noise corrupted signals of *Blocks, Bumps, HeaviSine and Doppler*

Proposition 5.2 *Let $f \in \mathbf{L}^2(\mathbb{R})$ be a piecewise smooth signal and $N(\cdot)$ denote a realization of a zero mean noise process. Suppose $\alpha_N(\cdot)$ characterize the pointwise Hölder exponent of $N(\cdot)$, then the measured data $Z = f + N$ admits $\alpha_N(\cdot)$ as its pointwise Hölder exponent except at the singular points of f.*

Proof:: see Appendix A

An ideal smoothness constrained filter, as noted earlier, is to isolate and localize the signal portion within the measured data. The following results establish that filtering by a multiscale curvature mask as defined in Equation (5.16) is equivalent to keeping the signal pointwise Hölder exponents at the singularity points of f and lifting its smoothness at all the remaining points.

Proposition 5.3 *Let $f, g \in \mathbf{L}^2(\mathbb{R})$ and $\psi(\cdot)$ be a compact support wavelet with n vanishing monents, and $Wf(\cdot, \cdot)$, $Wg(\cdot, \cdot)$ be the wavelet transforms of f and g respectively. Suppose*
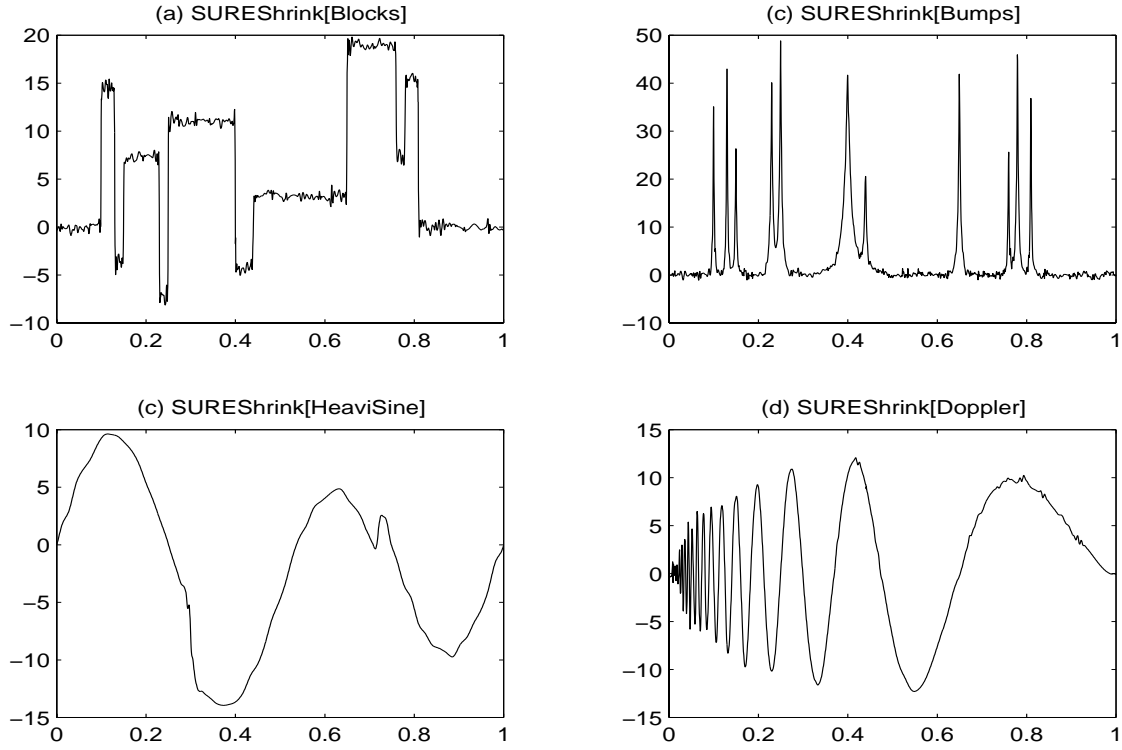
Figure 5.5: Denoising results by SureShrink for Gaussian noise corrupted signals

$\alpha_f(\cdot)$ and $\alpha_g(\cdot)$ are pointwise Hölder expontent functions of f, g and satisfy

$$\alpha_f(x) < n, \ \ \alpha_g(x) < n, \ \ \alpha_f(x) + \alpha_g(x) < n, \quad x \in \mathbb{R}$$

then

$$y = W^{-1}\{W f(s, u) W g(s, u)\}, \ \ (s, u) \in \mathbb{R}^+ \times \mathbb{R}$$

admits $\alpha_f(x) + \alpha_g(x) + \frac{1}{2}$ as its pointwise Hölder exponent for any $x \in \mathbb{R}$.

Proof:: see Appendix A

Proposition 5.4 *Let $\psi(\cdot)$ be a wavelet of compact support and with n vanishing monents. Consider the multiscale curvature mask as defined in Equation (5.15) and denote*

$$y = W^{-1}\{h(s, u)\}, \ \ (s, u) \in \mathbb{R}^+ \times \mathbb{R}$$

then y belongs to the Hölder space $C^n(\mathbb{R})$ almost everywhere.
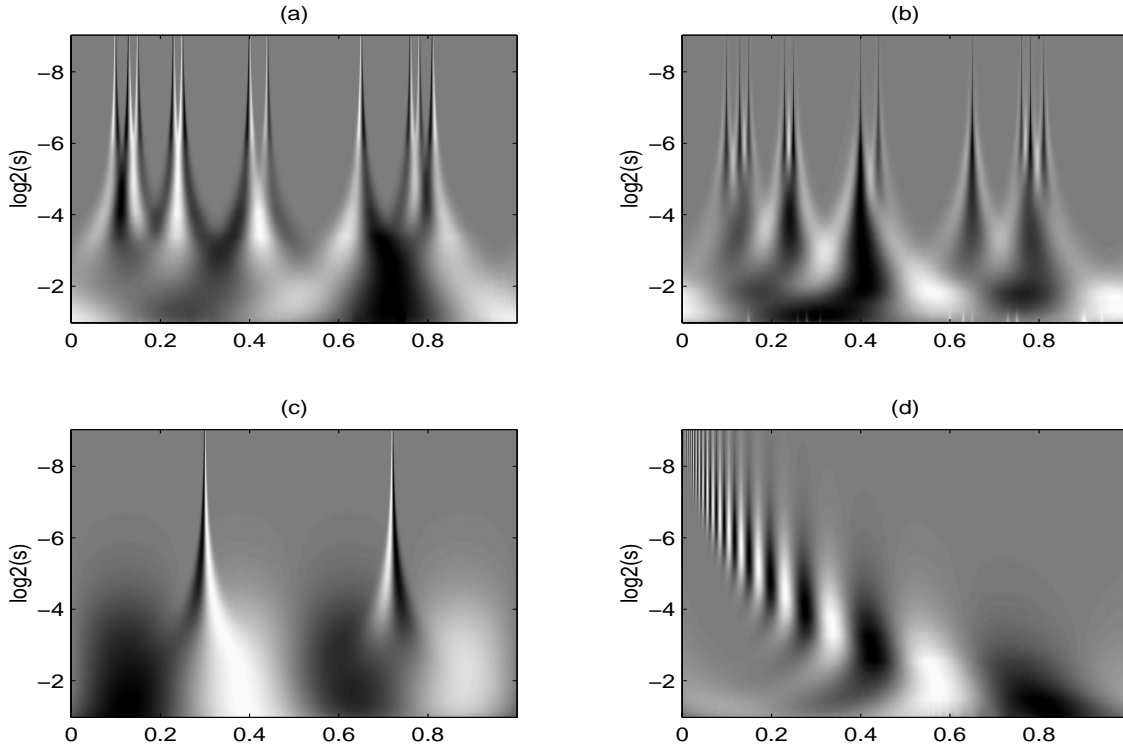
Figure 5.6: Multiscale curvature mask for (a) *Blocks*, (b) *Bumps*, (c) *HeaviSine* and (d) *Doppler*

Proof:: see Appendix A

Proposition 5.5 *Let $\psi(\cdot)$ be a wavelet of compact support and with n vanishing moments. Let $f \in \mathbf{L}^2(\mathbb{R})$ be a piecewise smooth signal. Its measured data $Z(\cdot)$ is modeled as in Equation (5.2). Suppose $\alpha_N(\cdot)$ characterizes the pointwise Hölder exponent of $N(\cdot)$. Let T be a smoothness constrained filter as defined in Equation (5.16) and $\hat{f} = TZ$. Then $V = Z - \hat{f}$ admits $\alpha_N(\cdot)$ as its pointwise Hölder exponent and $\hat{f}$ belongs to the Hölder space $C^n(I)$ except at the singular points of f where pointwise Hölder exponents of measured data $Z(\cdot)$ are preserved.*

Proof:: see Appendix A
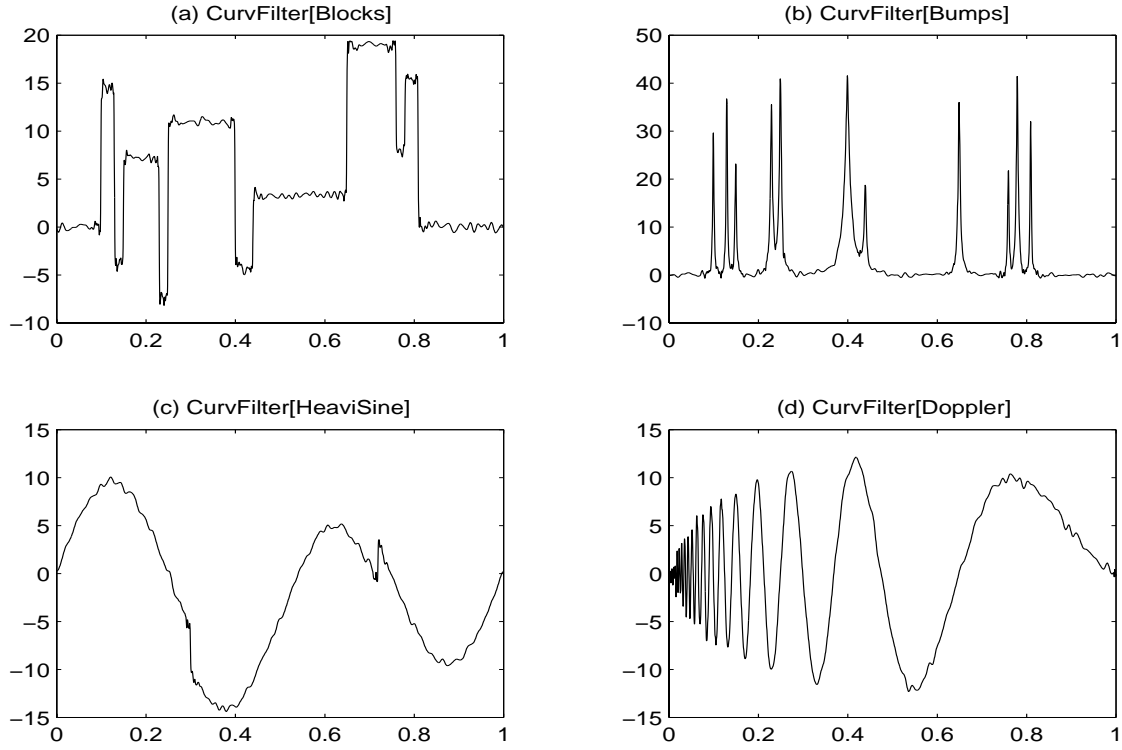
Figure 5.7: Filtering results by the smoothness constrained filter for Gaussian noise corrupted signals

5.6 Numerical Experiments

Four piecewise smooth signals, *Blocks, Bumps, HeaviSine and Doppler* are shown in Figure (5.3). These signals represent various spatially nonhomogeneous phenomena and are widely used in simulations of Wavelet Shrinkage methods.

 Figure (5.4) displays these signals corrupted by white Gaussian noise. The signal to noise ratio is $16.9dB$. Figure (5.5) shows the denoising results of Wavelet SureShrink [10] technique. By searching for the singular points of the signal from noisy data, we proceed to generate the multiscale curvature mask for the test signals as shown in Figure(5.6). Figure (5.7) demonstrates the results by the smoothness constrained filter as described in Section V. Compared to Wavelet SureShrink technique, our smoothness constrained filter preserves singularities of the underlying signals better and recovers the smoothness between singular points better. Filtering by a multiscale curvature
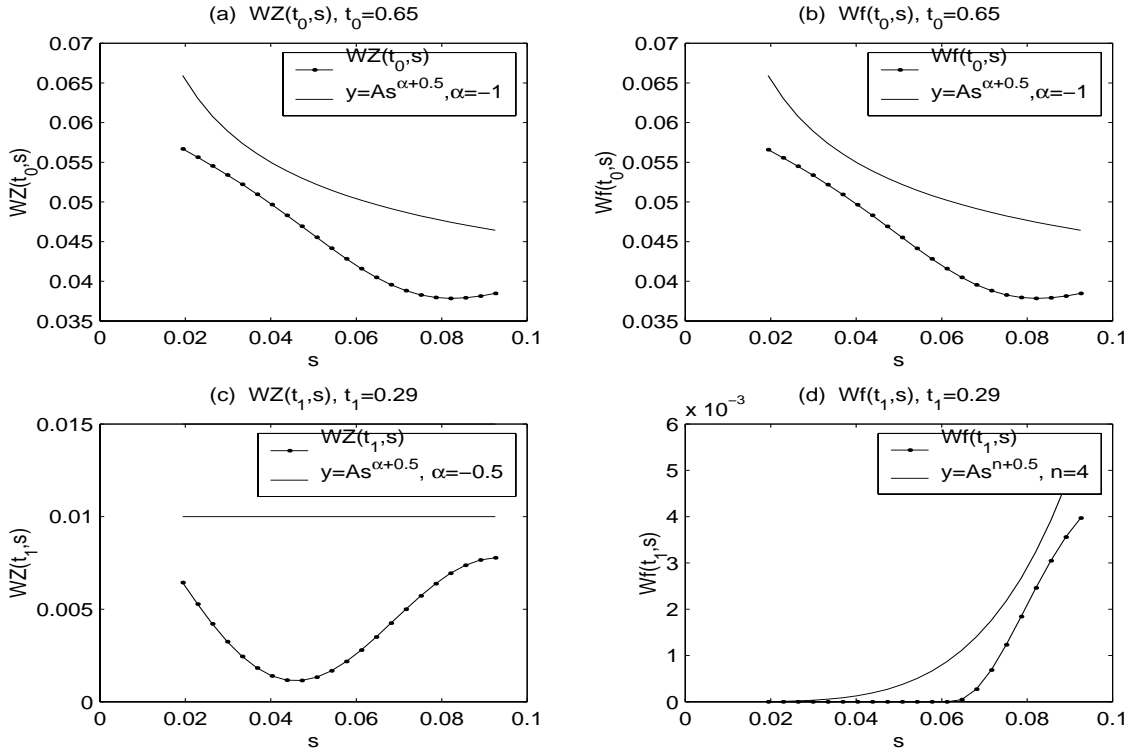
Figure 5.8: Adjustment of the wavelet transform amplitude decay across scales by smoothness constrained filter

mask is equivalent to keeping the signal pointwise Hölder exponents at the singular points of f, and lifting its smoothness at all the remaining points. This effect is demonstrated in Figure (5.8), which shows the adjustment of the wavelet transform amplitude decay of the noisy observation of the test signal *Blocks* at fine scales. As we can see in the filtered data, the wavelet transform amplitude decay at $t_0 = 0.65$, which corresponds to a singular point in *Blocks*, is preserved, and the decay at $t_0 = 0.29$, which is smooth in *Blocks*, is lifted.

In the non-Gaussian noise case, we test the above four signals which contaminated by Laplacian noise with SNR of $16.9dB$. Figure(5.9) shows the Laplacian corrupted noisy signals and Figure (5.10) demonstrates the denoising results from our proposed filter and Figure (5.11) shows the results from standard SureShrink techniques[1]. Laplacian

---

[1]Fully realizing SureShrink was not optimal for an non-Gaussian noise, this is used to merely illustrate our technique.
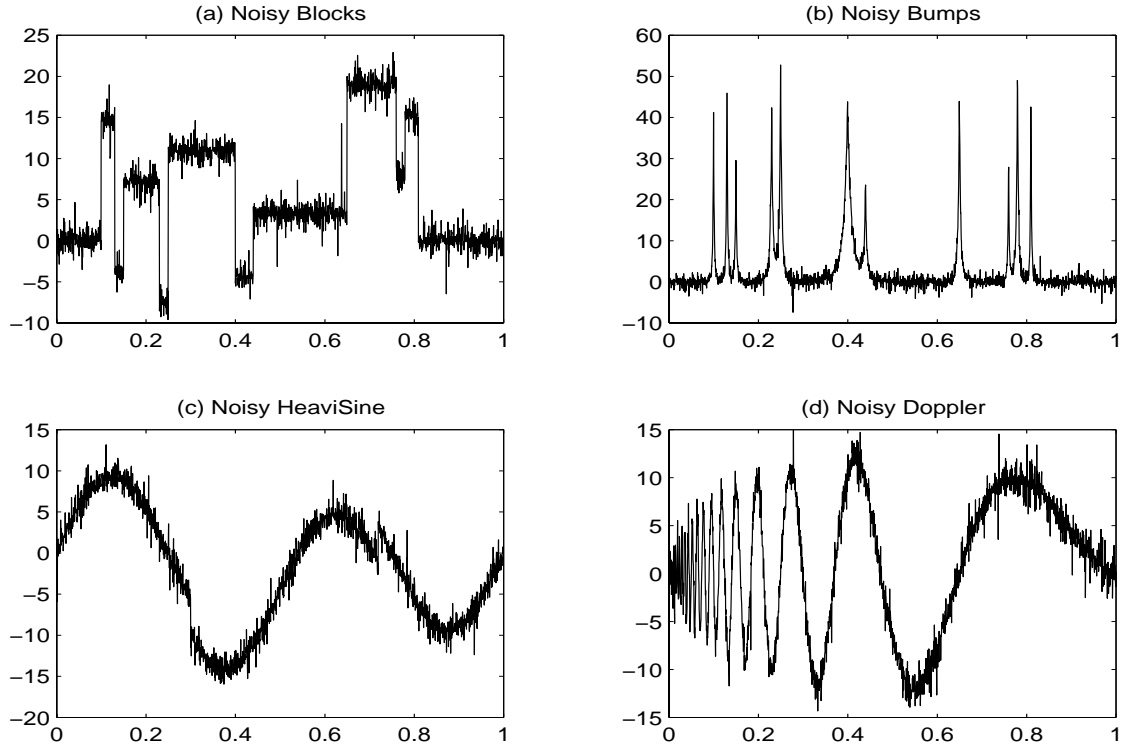
Figure 5.9: Laplacian noise corrupted signals of *Blocks, Bumps, HeaviSine and Doppler*

distribution has a heavier tail than Gaussian distribution which makes spurious spikes on noisy signals more likely. Wavelet Shrinkage methods are not robust to heavy tail noise, as we can see in the denoising result of SureShink, some spurious spikes are kept. However, performance of the proposed filter is not degraded as demonstrated in Figure (5.10).

Since the smoothness constrained filter preserves the decay of the wavelet transform amplitude at the singular point of f, it has the advantage of singularity preservation. In Figure (5.12), we test a planar shape, *B-52* contour, with the proposed filter to further demonstrate this desired property. Figure (5.12c) shows the result shape by our filter, in comparison with the result from a low pass Gaussian filter in Figure (5.12d), it is clear that the features of the shape is better preserved by the smoothness constrained filter.

It is important to note that our derivation does not rely on any statistical independence assumption, and can hence be extended to images. Suppose an image can be viewed as a differentiable function $z \in \mathbf{L}^2(\mathbb{R}^2)$, then we can generalize the curve evolution
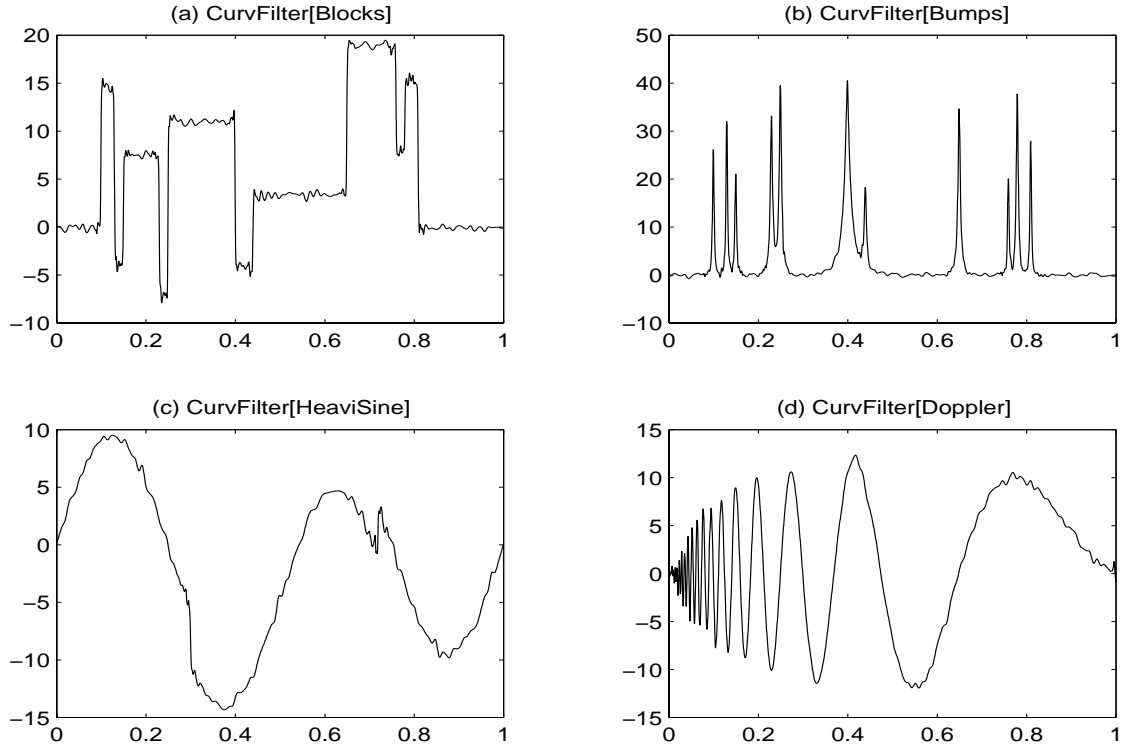
Figure 5.10: Filtering results by the smoothness constrained filter for Laplacian noise corrupted signals

equation (5.10) to two-dimensional case, $z(x, y, t) : \mathbb{R}^2 \times [0, T) \to \mathbb{R}$.

$$\begin{cases} \frac{\partial z(x,y,t)}{\partial t} = \frac{z_{xx} + z_{yy}}{1 + z_x^2 + z_y^2} = \frac{\Delta z}{1 + \|\nabla z\|^2} \\ z(x, y, 0) = z_0(x, y) \end{cases} \tag{5.17}$$

where z_0 is an initial smooth image in Cartesian coordinates. Applying the Maximum Principle, it follows directly that if $\|\nabla z\|^2$ is bounded, new extrema can not be created as z_0 evolves from fine to coarse scales. We therefore can find the true edges of z_0 by tracking the propagation line of its extrema across scale. Then, similarly to Section IV, we define $L(x, y), (x, y) \in \mathbb{R}^2$ to be the length of extrema propagation line originating at (x, y), and the set $\Omega = \{(x, y) : L(x, y) > T\}$ to represent the detected edge points. T is a threshold which we set to be half of the length of the longest extrema propagation line. Finally we define a edge indicator function $p(\cdot, \cdot)$ as

$$p(u_1, u_2) = \sum_{(x_i, y_i) \in \Omega} \delta(u_1 - x_i, u_2 - y_i). \tag{5.18}$$
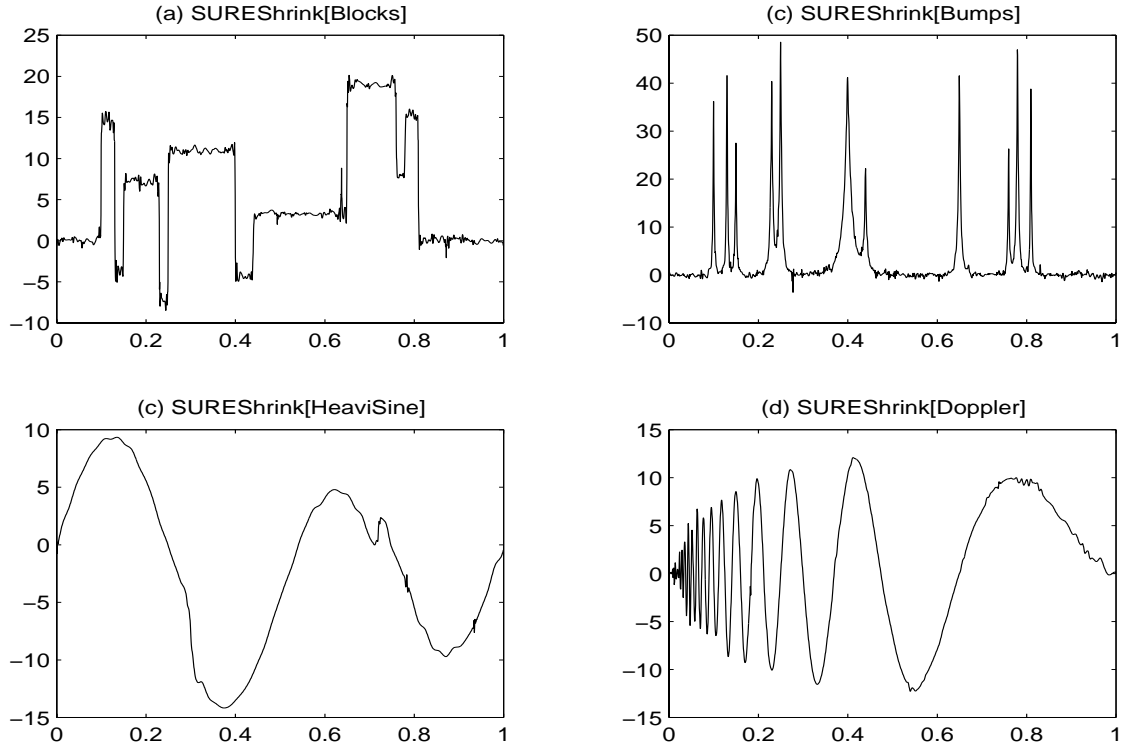
Figure 5.11: Denoising results by SureShrink for Laplacian noise corrupted signals

Consider an additive noise observation model, an image $f \in \mathbf{L}^2(\mathbb{R}^2)$ contaminated by addition of a noise. This noise is modeled by the realization of a zero mean random process $\{N\}$, i.e.,

$$Z(x,y) = f(x,y) + N(x,y), \quad (x,y) \in \mathbb{R}^2. \tag{5.19}$$

Let $\psi(\cdot)$ be a compactly supported wavelet with n vanishing moments. We define a two dimensional separable wavelet $\boldsymbol{\psi}(x,y) = \psi(x)\psi(y)$ and denote

$$\boldsymbol{\psi}_{s_1,u_1,s_2,u_2}(x,y) = \psi_{s_1,u_1}(x)\psi_{s_2,u_2}(y), \ s_1, s_2 > 0, (u_1, u_2) \in \mathbb{R}^2.$$

Let $WZ(s_1, u_1, s_2, u_2)$ denote the wavelet transformation of $Z(x,y)$, which is given by

$$WZ(s_1, u_1, s_2, u_2) = Z(x,y) \star \boldsymbol{\psi}_{s_1,u_1,s_2,u_2}(x,y). \tag{5.20}$$

By setting $z_0(x,y) = Z(x,y)$ in equation (5.17), an edge indicator function of the underlying image f can be obtained as in equation (5.18). With this edge indicator function,
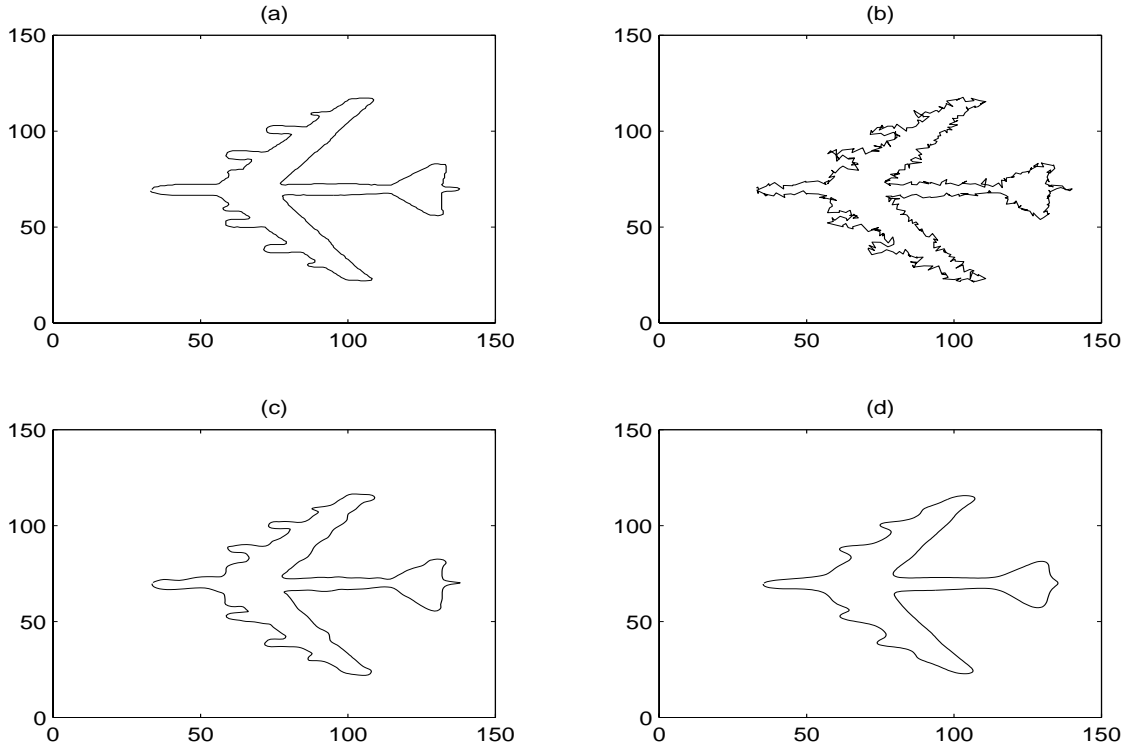
Figure 5.12: Denoising shapes by the smoothness constrained filter. (a) Shape *B-52* contour; (b) noise corrupted shape; (c) Resulting shape by the proposed filter; (d) Resulting shape by the low pass Gaussian filter.

we can proceed to generate a two-dimensional multiscale mask. Let

$$\rho(u_1, u_2) = \begin{cases} \exp\left(-\frac{u_1^2 + u_2^2}{1 - u_1^2 - u_2^2}\right), & u_1^2 + u_2^2 < 1 \\ 0, & u_1^2 + u_2^2 \geq 1 \end{cases}.$$

Then we define

$$q(s_1, u_1, s_2, u_2) = p(u_1, u_2) \star \rho\left(\frac{u_1}{C s_1}, \frac{u_2}{C s_2}\right), \quad s_1, s_2 > 0,$$

where C is the support of $\psi(\cdot)$. A two-dimensional multiscale mask can be given by

$$h(s_1, u_1, s_2, u_2) = \min\left(q(s_1, u_1, s_2, u_2),\ 1\right),$$

and the resulting filter for images may be defined as

$$\hat{f}(x, y) = \mathbf{W}_2^{-1}\{W Z(s_1, u_1, s_2, u_2) h(s_1, u_1, s_2, u_2)\}$$

where W_2^{-1} denotes a 2-D inverse wavelet transform.

Figure (5.13a) displays an original image *bar girl*, which is well known in the image processing literature. Figure (5.13b) shows the Gaussian noise corrupted image with a signal to noise ratio of $9.53dB$. By searching for the edge points of *bar girl* from noisy data as described in equation (5.17), an edge indicator function, displayed in Figure (5.13c), is obtained. We proceed to generate an two-dimensional mutliscale mask with this edge indicator function. Figure (5.13e) demonstrates the result by our proposed filter. The SNR is improved to nearly $3dB$. For non-Gaussian distributed noise, Figure (5.13f) displays a Laplacian noise corrupted image with a SNR of $9.54dB$. With the same procedure, the result by the proposed filter is shown in Figure (5.13f). In this case, the signal to noise ratio is increased to $12.44dB$. By the same argument for 1-D smoothness constrained filter, filtering by 2-D multiscale mask is equivalent to keeping image pointwise Hölder exponents at the edge points of the underlying image and lifting its smoothness at all the remaining points.

5.7 Conclusion

In this chapter, we proposed a novel non-linear smoothness constrained filtering technique. The key idea is to separate the signal portion out of its measured data, and to preserve the original smoothness property of the underlying signal. We first briefly reviewed the definition of Hölder space, Hölder exponent and established results of signal regularity measurement with wavelets. To detect the singular points of signal from measured data, we turned to curve shortening and derived the partial differential equations that characterize the evolution of curvature. A new singularity detection method by tracking the curvature extrema across scales is proposed and a multiscale curvature mask is generated. Then we proceed to project measured data into wavelet domain and suppress wavelet coefficients by this multiscale curvature mask. For a piecewise smooth signal, it can be shown that filtering by this curvature mask is equiv-

alent to keeping the signal pointwise Hölder exponents at the singular points of the underlying signal and lifting its smoothness at all the remaining points. An extension of the non-linear smoothness constrained filter to image processing is studied at the end of this paper. Numerical experiments demonstrated that our approach is effective and efficient.

5.8   Appendix

Proof of Proposition 5.2

Suppose $x \in \mathbb{R}$ is not a singular point.

Let $\alpha = \alpha_N(x)$, then there exists a polynomial $PN_x(\cdot)$ of degree $m = \lfloor \alpha \rfloor$ such that

$$\lim_{\Delta \to 0} \frac{|N(x + \Delta) - PN_x(\Delta)|}{|\Delta|^\beta} = 0, \ \ \beta < \alpha;$$

$$\lim_{\Delta \to 0} \frac{|N(x + \Delta) - PN_x(\Delta)|}{|\Delta|^\beta} = +\infty, \ \ \beta > \alpha.$$

Since $f(\cdot)$ is a piecewise smooth signal, f is infinitely differentiable at x and there exists a polynomial $Pf_x(\cdot)$ of degree $m = \lfloor \beta \rfloor$ for any $\beta > 0$ such that

$$\lim_{\Delta \to 0} \frac{|f(x + \Delta) - Pf_x(\Delta)|}{|\Delta|^\beta} = 0.$$

Let $PZ_x = PN_x + Pf_x$.

- For any $\beta < \alpha$,

$$\begin{aligned} &\lim_{\Delta \to 0} \frac{|Z(x + \Delta) - PZ_x(\Delta)|}{|\Delta|^\beta} \\ \leq \ &\lim_{\Delta \to 0} \frac{|f(x + \Delta) - Pf_x(\Delta)|}{|\Delta|^\beta} \\ + \ &\lim_{\Delta \to 0} \frac{|N(x + \Delta) - PN_x(\Delta)|}{|\Delta|^\beta} = 0. \end{aligned}$$

- For any $\beta > \alpha$,

$$\lim_{\Delta \to 0} \frac{|Z(x+\Delta) - PZ_x(\Delta)|}{|\Delta|^\beta}$$

$$\geq \lim_{\Delta \to 0} \frac{|N(x+\Delta) - PN_x(\Delta)|}{|\Delta|^\beta}$$

$$- \lim_{\Delta \to 0} \frac{|f(x+\Delta) - Pf_x(\Delta)|}{|\Delta|^\beta} = +\infty.$$

By the definition of Hölder exponent, we conclude that $Z(x)$ admits $\alpha_N(x)$ for all $x \in \mathbb{R}$ except the singular points of f. ∎

Proof of Proposition 5.3

f is Hölder $\alpha_f(x) < n$ at x, we have

$\forall (s, u) \in \mathbb{R}^+ \times \mathbb{R}$,

$$|Wf(s, u)| \leq A s^{\alpha_f(x) + \frac{1}{2}}\left(1 + \left|\frac{u - x}{s}\right|^{\alpha_f(x)}\right).$$

g is Hölder $\alpha_g(x) < n$ at x, we have

$\forall (s, u) \in \mathbb{R}^+ \times \mathbb{R}$,

$$|Wg(s, u)| \leq A s^{\alpha_g(x) + \frac{1}{2}}\left(1 + \left|\frac{u - x}{s}\right|^{\alpha_g(x)}\right).$$

therefore,

$$|Wf(s, u)Wg(s, u)| \leq A s^{\alpha_f(x) + \alpha_g(x) + 1}\left(1 + \right.$$
$$\left.\left|\frac{u - x}{s}\right|^{\alpha_f(x)} + \left|\frac{u - x}{s}\right|^{\alpha_g(x)} + \left|\frac{u - x}{s}\right|^{\alpha_f(x) + \alpha_g(x)}\right).$$

Let

$$\left|\frac{u - x}{s}\right|^{\alpha'} = \max\left(\left|\frac{u - x}{s}\right|^{\alpha_f(x)}, \right.$$
$$\left. \left|\frac{u - x}{s}\right|^{\alpha_g(x)}, \left|\frac{u - x}{s}\right|^{\alpha_f(x) + \alpha_g(x)}\right)$$

and it is then clear that $\alpha' < \alpha_f(x) + \alpha_g(x) + \frac{1}{2}$. Then we can rewrite the above equation as

$$|Wf(s, u)Wg(s, u)| \leq A s^{\alpha_f(x) + \alpha_g(x) + 1}\left(1 + \left|\frac{u - x}{s}\right|^{\alpha'}\right).$$

by *theorem 5.1* we conclude that

$$y = \mathbf{W}^{-1}\{Wf(s, u)Wg(s, u)\}$$

admits $\alpha_f(x) + \alpha_g(x) + \frac{1}{2}$ as its pointwise Hölder exponent for any $x \in \mathbb{R}$. ∎

Proof of Proposition 5.4

We first introduce without proof a theorem from Meyer [16] and Mallat [70].

Theorem 5.2 *Let $\psi(\cdot)$ be a wavelet with n vanishing moments and $f \in \mathrm{L}^2(\mathbb{R})$ belongs to a global Hölder space $C^\alpha(a, b)$, and $Wf(\cdot, \cdot)$ denotes its wavelet transform. Then there exists a constant $A > 0$ such that*

$$\forall (s, u) \in \mathbb{R}^+ \times (a, b), |Wf(s, u)| \leq As^{\alpha + \frac{1}{2}}. \tag{5.21}$$

Conversely, suppose f is bounded and $Wf(s, u)$ satisfies equation (5.21) for an $\alpha < n$ that is not an integer. Then f belongs to the global Hölder space $C^\alpha(a + \epsilon, b - \epsilon)$ for any $\epsilon > 0$.

Let $\Omega = \{x_i\}_{i=1}^{N-1}$ be the set of all the singular points of f. For convenience, denote $x_0 = -\infty$ and $x_N = +\infty$. Observe that for any $\epsilon > 0$, there exists s_0 such that

$$h(s, u) = 0, \quad \forall s < s_0, \; u \in (x_i + \epsilon, x_{i+1} - \epsilon),$$

$i = 0, 1, ...N - 1$. Therefore, by an appropriate choice of A, we obtain

$$|h(s, u_0)| \leq As^{n + \frac{1}{2}}, \quad \forall (s, u) \in \mathbb{R}^+ \times (x_i + \epsilon, x_{i+1} - \epsilon),$$

$i = 0, 1, ...N - 1$. This concludes the proof of *proposition 5.4*. ∎

Proof of Proposition 5.5

Let $\Omega = \{x_i\}_{i=1}^{N-1}$ be the set of all the singular points of f and denote $x_0 = -\infty$ and $x_N = +\infty$. By *propositions 5.3* and *5.4*, it follows directly that $\hat{f}$ belongs to the global Hölder space $C^n(x_i + \epsilon, x_{i+1} - \epsilon)$ for any $\epsilon > 0, \; i = 0, 1, ...N - 1$. Let $v = x_i$ and suppose the pointwise Hölder exponent of Z at v is α, then $\forall (s, u) \in \mathbb{R}^+ \times \mathbb{R}$,

$$\begin{aligned} |\hat{W}f(s, u)| &= |WZ(s, u)h(s, u)| \\ &\leq As^{\alpha + \frac{1}{2}}(1 + |\frac{u - v}{s}|^{\alpha'}). \end{aligned}$$

Hence, the pointwise Hölder exponents of measured data $Z(\cdot)$ are preserved at the singular points of f. Let $V = Z - \hat{f}$. Observe for any $\epsilon > 0$, there exists s_0 such that for all $s < s_0$,

$$u \in (x_i + \epsilon, x_{i+1} - \epsilon), \quad WV(s, u) = WZ(s, u),$$

and for all $s \geq s_0$,

$$u \in (x_i + \epsilon, x_{i+1} - \epsilon), \quad WV(s, u) \leq WZ(s, u),$$

$i = 0, 1, ...N - 1$. By *proposition 5.2* and *theorem 5.1*, $V(\cdot)$ admits $\alpha_N(\cdot)$ as its pointwise Hölder exponent except at the singular points of f. This concludes the proof of *proposition 5.5*. ∎

Figure 5.13: (a) Original *bar girl*; (b) *bar girl* plus Gaussian noise; (c) Edge points extracted from noisy image; (d) Filtered result from the Gaussian noise corrupted image; (e) *bar girl* plus Laplacian noise; (f) Filtered result from the Laplacian noise corrupted image;

6 A Multiscale Approach for Pixel-Level Image Fusion

6.1 Introduction

WITH the development of new imaging sensors arises the need for image processing techniques that can effectively fuse images from different sensors into a single composite entity for interpretation. Fusion begins with two or more registered images that contain different representations of the same scene. They may come from different viewing conditions, or even different sensors (visible and IR, CT and MRI, etc.). Image fusion of multiple sensors in a vision system could significantly reduce human/machine error in detection and recognition of objects due to the inherent redundancy and extended coverage. For example, fusion of forward looking infrared (FLIR) and low light television images (LLTV) obtained by an airborne sensor platform would aid a pilot to navigate in poor weather conditions.

The actual fusion process can take place at different levels of information representation. A generic categorization is to consider the different processes at signal, pixel, feature and symbolic levels [23]. We focus on the so-called pixel level fusion process, where a composite image has to be constructed from several input images. Some
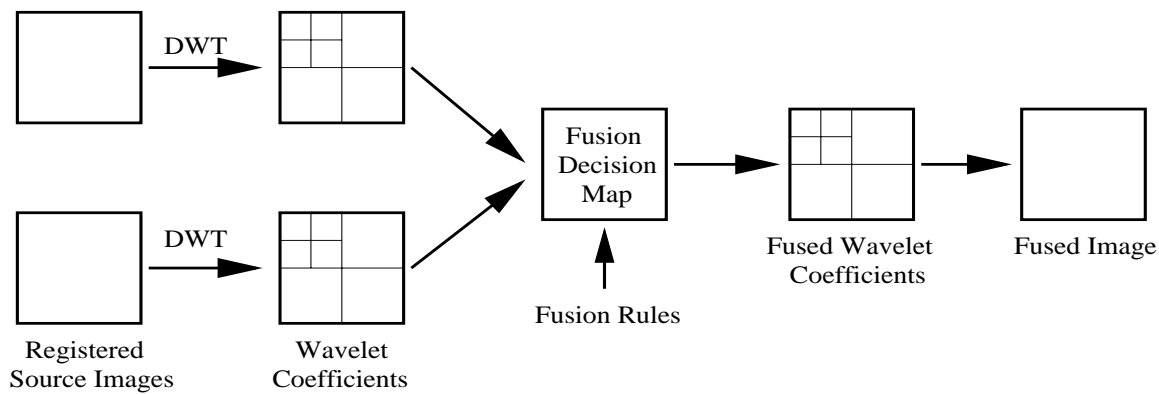
Figure 6.1: A general framework for multiscale fusion with wavelet transform

generic requirements must be imposed on the fusion result. The fusion process should preserve all relevant information of the input imagery in the composite image (pattern conservation). The fusion scheme should not introduce any artifacts or inconsistencies which would distract a human observer or the following processing stages (Causality). Over the past two decades, a wide variety of pixel-level image fusion algorithms has been developed. These techniques may be classified into linear superposition, logical filter [17], mathematical morphology [18], image algebra [19] [20], artificial neural network [21], and simulated annealing [22] methods. Each of these algorithms focuses on the fact that the fused image reveals new information concerning features that can not be perceived in individual sensor images. However, some useful information has been discarded since each fusion scheme tends to emphasize different attributes of the image. Luo [23] provides a detailed review of these techniques.

 Inspired by the fact that the human visual system processes and analyzes image information at different scales, researchers recently proposed a multiscale based fusion method which is widely accepted [24] as one of the most effective techniques for image fusion. A multicale transform, which may be a pyramid or wavelet transform, is first calculated for each input image, then a composite is formed by selecting the coefficients from the multicale representations of all the source images. Finally, a fused image is recovered through an inverse transformation. In the pioneering work of Burt [28], a Laplacian pyramid and a "choose max" rule is proposed as a model for binocular fu-

sion in human stereo vision. For each coefficient in the pyramids of source images, the one with the maximum amplitude is copied to the composite pyramid and the fused image is reconstructed from an inverse pyramid transform. More recently [95], fusion within a gradient pyramid was shown to provide improved stability and noise immunity.

Wavelet theory has played a particularly important role in multiscale analysis. A number of papers [25] [26] [27] have addressed fusion algorithms based on the orthogonal wavelet transform. A general framework for multiscale fusion with wavelet transform is shown in Figure (6.1). The wavelet transform offers certain advantages over the Laplacian pyramid-based techniques. Since the wavelet bases are chosen to be orthogonal, the information gleaned at different resolutions is unique; on the other hand, the pyramid decomposition contains redundancy between different scales. Furthermore, a wavelet image representation provides directional information in the high-low, low-high and high-high bands, while the pyramid representation fails to introduce any spatial orientation selectivity into the decomposition process. A major drawback in the recent pursuit of wavelet-based fusion algorithms is due to a lack of a good fusion scheme. Most selection rules so far proposed are in essence more or less similar to "choose max", which introduces a significant amount of high frequency noise due to the sudden switch of the fused wavelet coefficient to that which is maximum of the source. This high frequency noise is particularly undesirable to visual perception.

In this chapter, we apply a biorthogonal wavelet transform to the pixel level image fusion. It is possible to construct smooth biorthogonal wavelets of compact support which are either symmetric or antisymmetric. At the exception of a Haar wavelet, it has been shown [29] that symmetric orthogonal wavelets are impossible to construct. Symmetric or antisymmetric wavelets are synthesized with perfect reconstruction filters having a linear phase. This is a desirable property for image fusion applications. Unlike the "choose max" type of selection rules, we propose an information theoretic fusion scheme. For each pixel in a source image, a vector consisting of wavelet coefficients at that pixel position across scales is formed to indicate the "activity" of that

pixel. We denote these indicator vectors of all the pixels in a source image as its activity map. A decision map is then obtained by applying an information theoretic divergence to measure all the source activity maps. To make a reasonable comparison among activity indicator vectors, we apply our newly proposed divergence measure, Jensen-Rényi divergence, which is defined in terms of Rényi entropy [5]. Wavelet coefficients of the fused image are finally selected according to the decision map. Since all the scales, from fine to coarse, are considered to evaluate the activity at a particular position within an image, our approach is intrinsically more accurate, in the sense of selecting coefficients containing the richest information, relative to the "select max" type of fusion schemes. This chapter is organized as follows. We briefly introduce a biorthognal wavelet representation of images in Section II. A concise formulation of the problem is given in Section III. In Section IV, we describe the information theoretic fusion algorithm and present some numerical experiments. Finally, we provide concluding remarks in Section V.

6.2   Biorthogonal Wavelet Representation of Images

Let $\{\phi, \tilde{\phi}\}$ and $\{\psi, \tilde{\psi}\}$ be two dual pairs of scaling functions and wavelets which generate biorthogonal wavelet basis of $\mathbf{L}^2(\mathbb{R})$. For $\boldsymbol{t} = (t_1, t_2)$ and $\boldsymbol{n} = (n_1, n_2)$, we write

$$\phi^2_{j,\boldsymbol{n}}(\boldsymbol{t}) = \phi_{j,n_1}(t_1)\phi_{j,n_2}(t_2), \tag{6.1}$$

$$\psi^1(\boldsymbol{t}) = \phi(t_1)\psi(t_2), \quad \psi^2(\boldsymbol{t}) = \psi(t_1)\phi(t_2), \quad \psi^3(\boldsymbol{t}) = \psi(t_1)\psi(t_2), \tag{6.2}$$

$$\tilde{\psi}^1(\boldsymbol{t}) = \tilde{\phi}(t_1)\tilde{\psi}(t_2), \quad \tilde{\psi}^2(\boldsymbol{t}) = \tilde{\psi}(t_1)\tilde{\phi}(t_2), \quad \tilde{\psi}^3(\boldsymbol{t}) = \tilde{\psi}(t_1)\tilde{\psi}(t_2). \tag{6.3}$$

Denote for $1 \le k \le 3$,

$$\psi^k_{j,\boldsymbol{n}}(\boldsymbol{t}) = \frac{1}{2^j}\psi^k\left(\frac{t_1 - 2^j n_1}{2^j}, \frac{t_2 - 2^j n_2}{2^j}\right) \tag{6.4}$$

and

$$\tilde{\psi}^k_{j,\boldsymbol{n}}(\boldsymbol{t}) = \frac{1}{2^j}\tilde{\psi}^k\left(\frac{t_1 - 2^j n_1}{2^j}, \frac{t_2 - 2^j n_2}{2^j}\right). \tag{6.5}$$

It is easy to verify [70] that

$$\{\psi^1_{j,\boldsymbol{n}}, \psi^2_{j,\boldsymbol{n}}, \psi^3_{j,\boldsymbol{n}}\}_{(j,\boldsymbol{n})\in\mathbb{Z}^3}$$

and

$$\{\tilde{\psi}^1_{j,\boldsymbol{n}}, \tilde{\psi}^2_{j,\boldsymbol{n}}, \tilde{\psi}^3_{j,\boldsymbol{n}}\}_{(j,\boldsymbol{n})\in\mathbb{Z}^3}$$

are biorthogonal bases of $\mathbf{L}^2(\mathbb{R}^2)$.

Any $f \in \mathbf{L}^2(\mathbb{R}^2)$ has two possible decompositions in these bases,

$$f = \sum_j \sum_{\boldsymbol{n}} \sum_{k=1}^3 < f, \psi^k_{j,\boldsymbol{n}} > \tilde{\psi}^k_{j,\boldsymbol{n}} = \sum_j \sum_{\boldsymbol{n}} \sum_{k=1}^3 < f, \tilde{\psi}^k_{j,\boldsymbol{n}} > \psi^k_{j,\boldsymbol{n}}, \tag{6.6}$$

where $< \cdot, \cdot >$ denotes an inner product of two functions. Assuming we choose $\psi^k_{j,\boldsymbol{n}}$ as the analysis wavelets, at any scale 2^j, we denote the approximation coefficient

$$a_j[\boldsymbol{n}] = < f, \phi^2_{j,\boldsymbol{n}} >$$

and the wavelet coefficient

$$d^k_j[\boldsymbol{n}] = < f, \psi^k_{j,\boldsymbol{n}} >, \;\; k = 1, 2, 3.$$

The three wavelets ψ^k extract image details at different scales and orientations. Over positive frequencies, $\hat{\phi}$ and $\hat{\psi}$ have an energy mainly concentrated respectively on lower and higher frequencies. Let $\boldsymbol{\omega} = (\omega_1, \omega_2)$, the separable wavelet expressions implies that

$$\hat{\psi}^1(\boldsymbol{\omega}) = \hat{\phi}(\omega_1)\hat{\psi}(\omega_2), \;\; \hat{\psi}^2(\boldsymbol{\omega}) = \hat{\psi}(\omega_1)\hat{\phi}(\omega_2), \;\; \hat{\psi}^3(\boldsymbol{\omega}) = \hat{\psi}(\omega_1)\hat{\psi}(\omega_2). \tag{6.7}$$

Hence $|\hat{\psi}^1(\boldsymbol{\omega})|$ emphasizes low horizontal frequencies ω_1 and high vertical frequencies ω_2, while $|\hat{\psi}^2(\boldsymbol{\omega})|$ is larger at high horizontal frequencies ω_1 and low vertical frequencies ω_2, whereas $|\hat{\psi}^3(\boldsymbol{\omega})|$ is larger at both high horizontal frequencies ω_1 and high vertical frequencies ω_2. As a result, wavelet coefficients calculated with ψ^1 and ψ^2 are larger along edges which are respectively horizontal and vertical, and ψ^3 produces large coefficients at the corners.

The wavelet coefficients at scale 2^{j+1} are calculated from a_j with two dimensional separable convolutions and subsamplings. Let $\{h, g\}$ and $\{\tilde{h}, \tilde{g}\}$ be the perfect reconstruction filters associated to the biorthogonal wavelet $\{\psi, \tilde{\psi}\}$. For any pair of one-dimensional
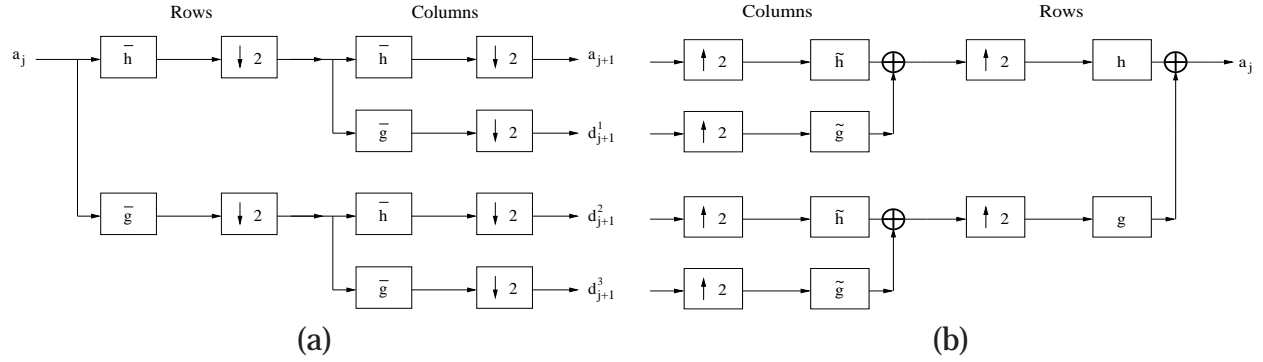
Figure 6.2: A fast biorthogonal two-dimensional wavelet transform (a) and its inverse transform (b) implemented by perfect reconstruction filter banks.

filters $y[n]$ and $z[n]$, we write the product filter $yz[\boldsymbol{n}] = y[n_1]z[n_2]$, and denote $\bar{y}[n] = y[-n]$. For $\boldsymbol{n} = (n_1, n_2)$,

$$a_{j+1}[\boldsymbol{n}] = a_j \star \bar{h}\bar{h}[\boldsymbol{n}], \tag{6.8}$$

$$d^1_{j+1}[\boldsymbol{n}] = a_j \star \bar{h}\bar{g}[\boldsymbol{n}], \tag{6.9}$$

$$d^2_{j+1}[\boldsymbol{n}] = a_j \star \bar{g}\bar{h}[\boldsymbol{n}], \tag{6.10}$$

$$d^3_{j+1}[\boldsymbol{n}] = a_j \star \bar{g}\bar{g}[\boldsymbol{n}], \tag{6.11}$$

where "$\star$" denotes convolution. A separable two dimensional convolution can be factored into one-dimensional convolutions along with rows and columns of the image. The factorization is illustrated in Figure (6.2a). The rows of a_j are first convolved with $\bar{h}$ and $\bar{g}$, and subsampled by 2. The columns of these two output images are then convolved respectively with $\bar{h}$ and $\bar{g}$ and subsampled, which yields four subsampled images a_{j+1}, d^1_{j+1}, d^2_{j+1} and d^3_{j+1}.

We denote

$$\breve{y}[\boldsymbol{n}] = \breve{y}[n_1, n_2] = \begin{cases} y[k_1, k_2] & , \ n_1 = 2k_1, n_2 = 2k_2 \\ 0 & , \ \text{otherwise} \end{cases},$$

the image obtained by inserting a row of zeros and a column of zeros between pairs of consecutive rows and columns of $y[n_1, n_2]$. a_j is recovered from the coarser scale approximation a_{j+1} and the wavelet coefficients d^1_{j+1}, d^2_{j+1} and d^3_{j+1} with two-dimensional

separable convolutions

$$a_j[\boldsymbol{n}] = \breve{a}_{j+1} \star \tilde{h}\tilde{h}[\boldsymbol{n}] + \breve{d}_{j+1}^1 \star \tilde{h}\tilde{g}[\boldsymbol{n}] + \breve{d}_{j+1}^2 \star \tilde{g}\tilde{h}[\boldsymbol{n}] + \breve{d}_{j+1}^3 \star \tilde{g}\tilde{g}[\boldsymbol{n}]. \qquad (6.13)$$

These four convolutions can also be factored into six groups of one-dimensional convolutions along rows and columns, as illustrated in Figure (6.2b).

Let $a_J[\boldsymbol{n}]$ be a digital image whose pixel interval equals to $2^J = N^{-1}$. A biorthogonal wavelet image representation of a_J of depth $L - J$ is computed by iterating Equation (6.8-6.11) for $J < j \leq L$:

$$\{d_j^1, d_j^2, d_j^3, a_L\}_{J < j \leq L}. \qquad (6.14)$$

The original digital image a_J is recovered from this wavelet representation by iterating the reconstruction Equation (6.13) for $J < j \leq L$.

6.3 Problem Formulation

Let $f_1, f_2, ..., f_n : \mathbb{Z}^2 \to \mathbb{R}$ be digital images of the same scene taken from different sensors. For the pixel level image fusion problem, we assume all the source images are registered so that the difference in resolution, coverage, treatment of a theme, characteristics of the image acquisition methods are eliminated. The goal of our fusion algorithm is to construct a composite image such that information captured in all the source images are combined and the source image data is hence compressed. To achieve this goal, we apply an information theoretic fusion approach based on a biorthogonal wavelet representation.

Definition 6.1 *Let $W f_i = \{d_i^1(j, \boldsymbol{n}), d_i^2(j, \boldsymbol{n}), d_i^3(j, \boldsymbol{n}), a_i(L, \boldsymbol{n})\}_{0 < j \leq L, \boldsymbol{n} \in \mathbb{Z}^2}, \ i = 1, 2, ...n$ be a biorthogonal wavelet image representation of f_i as defined in equation (6.14). With no loss of generality, we set $J = 0$. For any $\boldsymbol{n} \in \mathbb{Z}^2$, an activity pattern vector is defined as*

$$A_i(\boldsymbol{n}) = \left[\sum_{k=1}^{3} |d_i^k(1, 2^{L-1}\boldsymbol{n})|^2, \sum_{k=1}^{3} |d_i^k(2, 2^{L-2}\boldsymbol{n})|^2, ..., \sum_{k=1}^{3} |d_i^k(L, \boldsymbol{n})|^2 \right], \qquad (6.15)$$

which is a $(1 \times L)$ *vector of energy concentrated at pixel* $f_i(2^L \boldsymbol{n})$ *across scales. We refer to* $\{A_i(\boldsymbol{n})\}_{\boldsymbol{n} \in \mathbb{Z}^2}$ *as the activity map of source image* f_i, $i = 1, 2, ...n$.

Activity maps characterize the inherent information pattern in source images. To fuse the source wavelet coefficients, it is necessary to compare the activity patterns for every pixel. For instance, if the activity patterns are different in some region, taking the average of wavelet coefficients to generate a composite image is not a good choice, since it would create artifacts. On the other hand, if the activity patterns are similar for that region, taking the average would inject more information to the fused image due to the contribution from different sources.

A reasonable measure for activity patterns should satisfy the following properties:

- It should be capable of measuring the difference between two or more activity patterns.

- It should be nonnegative and symmetric.

- It should vanish to zero if and only if the activity patterns are the same.

- It should reach the maximum value when activity patterns are degenerate distributions.

In Chapter 3, we proposed a new information divergence measure, referred to as Jensen-Rényi divergence, which satisfies the above requirements. A decision map is then generated by applying the Jensen-Rényi divergence to measure the coherence of source activity maps at the pixel level. We further segment the decision map into two regions, D_0 and D_1. D_0 is the set of pixels whose activity patterns are similar in all the source images, while D_1 is the set of pixels whose activity patterns are different. Our fusion scheme is to find the solution to the following optimization problem.

$$Wf = \arg \min_{Wf \in \mathcal{F}} \left\{ \sum_{i=1}^{n} \left(\sum_{j=1}^{L} \sum_{2^j \boldsymbol{n} \in D_0} |Wf(j, \boldsymbol{n}) - Wf_i(j, \boldsymbol{n})|^2 \right) - \sum_{j=1}^{L} \sum_{2^j \boldsymbol{n} \in D_1} |Wf(j, \boldsymbol{n})|^2 \right\},$$

(6.16)

where $\mathcal{F}$ is the set of all the images f whose wavelet transform satisfies

$$\min_i(Wf_i(j, \boldsymbol{n})) \leq Wf(j, \boldsymbol{n}) \leq \max_i(Wf_i(j, \boldsymbol{n})),$$

for any $0 < j \leq L$ and $\boldsymbol{n} \in \mathbb{Z}^2$. This constraint makes sure that the solution stays in the closure of $\mathcal{F}$, i.e., no image outside the scenario we are contemplating.

6.4 Image Fusion with Jensen-Rényi divergence

The goal of image fusion is to integrate complementary information from multi-sensor data such that the fused images are more suitable for the purpose of human visual perception. Let $f_1, f_2, ... f_n : \mathbb{Z}^2 \to \mathbb{R}$ be the digital images generated by different sensors. Our information theoretic fusion approach first calculates a biorthogonal wavelet image representation for each f_i, then a pixel level activity map $\{A_i(\boldsymbol{n})\}_{\boldsymbol{n} \in \mathbb{Z}^2}$ is formed, as described in Section III. Denote $\| \cdot \|$ as l_1 norm, for any $\boldsymbol{n} \in \mathbb{Z}^2$, we define a normalized activity pattern

$$\boldsymbol{p}_i(\boldsymbol{n}) = \begin{cases} A_i(\boldsymbol{n})/\|A_i(\boldsymbol{n})\| & , \|A_i(\boldsymbol{n})\| \neq 0 \\ \Delta_1 & , \|A_i(\boldsymbol{n})\| = 0 \end{cases}$$

where $\Delta_1 = [1, 0, ...0]$ is a $(1 \times L)$ degenerate distribution. To fuse the source wavelet coefficients, we compare the normalized activity patterns of all the source images in terms of Jensen-Rényi divergence, and create a selection map $\{S(\boldsymbol{n})\}_{\boldsymbol{n} \in \mathbb{Z}^2}$:

$$S(\boldsymbol{n}) = JR_\alpha^{\boldsymbol{\omega}}(\boldsymbol{p}_1(\boldsymbol{n}), ..., \boldsymbol{p}_n(\boldsymbol{n})). \tag{6.17}$$

The selection map is further segmented into two decision regions, D_0 and D_1. Set T to be the mean value of selection map, we write

$$D_0 = \{\boldsymbol{n} \in \mathbb{Z}^2 : S(\lfloor 2^{-L}\boldsymbol{n} \rfloor) < T\}$$

and

$$D_1 = \{\boldsymbol{n} \in \mathbb{Z}^2 : S(\lfloor 2^{-L}\boldsymbol{n} \rfloor) \geq T\},$$

where $\lfloor x \rfloor$ denotes the integer part of x.

Since $D_0 \cap D_1 = \emptyset$, the solution to Equation (6.16) can be obtained by searching for each wavelet coefficient of the fused image individually. Let f be a composite image defined by its wavelet coefficients

$$Wf = \{d_f^1(j, \boldsymbol{n}), d_F^2(j, \boldsymbol{n}), d_f^3(j, \boldsymbol{n}), a_f(L, \boldsymbol{n})\}_{0 < j \le L, \boldsymbol{n} \in \mathbb{Z}^2},$$

where

$$a_f^k(L, \boldsymbol{n}) = \begin{cases} (1/n) \sum_{i=1}^n a_i^k(L, \boldsymbol{n}), & 2^L \boldsymbol{n} \in D_0 \\ \max(a_i^k(L, \boldsymbol{n})), & 2^L \boldsymbol{n} \in D_1 \end{cases},$$

and for $k = 1, 2, 3$,

$$d_f^k(j, \boldsymbol{n}) = \begin{cases} (1/n) \sum_{i=1}^n d_i^k(j, \boldsymbol{n}), & 2^j \boldsymbol{n} \in D_0 \\ \max(d_i^k(j, \boldsymbol{n})), & 2^j \boldsymbol{n} \in D_1 \end{cases}.$$

It is thus easy to verify that f satisfies our fusion criteria (6.16).

Four examples, including multi-sensor navigation image fusion, multi-modality medical image fusion, multi-spectral remote sensing image fusion, and multi-focus optical image fusion are now presented to illustrate the fusion scheme defined above.

6.4.1   Multi-Sensor Image Fusion

To help helicopter pilots navigate under poor visibility conditions, such as fog or heavy rain, helicopters are equipped with several imaging sensors, which can be viewed by the pilot in a helmet mounted display. A typical sensor suite includes both a low-light-television (LLTV) sensor and a thermal imaging forward-looking-infrared (FLIR) sensor. In the current configuration, the pilot can choose one of the two sensors to watch in his display. Sample LLTV and FLIR images are shown in Figure (6.3.1) and (6.3.2) respectively. A possible improvement is to combine both imaging sources into a single fused image.

Image fusion by standard techniques such as pixel averaging and multiscale based maximum selection scheme are shown in Figure (6.3.3) and (6.3.4) respectively. Note that the pixel averaging method has a muddy appearance. This is due primarily to the

fact that averaging results in reduced contrast for all the patterns which appear in only one source. On the other hand, maximum selection scheme produces some mosaic like artifacts due to the high frequency noise introduced by a sudden switch between two sets of source wavelet coefficients. Image fusion with our multiscale information theoretic approach is illustrated in Figure (6.3.6). As we can see, all the significant features from both sources are retained in the fused image without suffering from artifacts.

6.4.2 Multi-Modality Image Fusion

With the development of new imaging methods in medical diagnostics arises the need of meaningful and spatial correct combination of all available image datasets. Examples for imaging devices include computer tomography (CT), magnetic resonance imaging (MRI) or the newer positron emission tomography (PET). Image fusion of a CT (Figure 6.4.1) and a MRI (Figure 6.4.2) image with our multiscale information theoretic approach is illustrated in Figure (6.4.6). For comparison purpose, fusion by pixel averaging and by multiscale based maximum selection scheme are shown in Figures (6.4.3) and (6.4.4).

6.4.3 Multi-Spectral Image Fusion

Image fusion is often involved in remote sensing: modern spectral sensors gather up to several hundreds of spectral bands which can be either visualized and processed individually, or can be fused into a single image, depending on the image analysis task. Image fusion of two bands from a multispectral sensor with our multiscale information theoretic approach is illustrated in Figure (6.5). For comparison purpose, fusion by pixel averaging and by multiscale based maximum selection scheme are shown in Figures (6.5.3) and (6.5.4).

6.4.4   Multi-Focus Image Fusion

Due to the limited depth-of-focus of optical lenses, it is often impossible to get an image
which contains all relevant objects 'in focus'. One possibility to overcome this problem
is to take several pictures with different focus points and combine them together into
a single frame which finally contains the focused regions of all input images. Figure
(6.6) illustrates our multiscale information theoretic fusion approach. For comparison
purpose, fusion by pixel averaging and by multiscale based maximum selection scheme
are shown in Figures (6.6.3) and (6.6.4).

6.4.5   Performance Measures

It is difficult to define a general performance measure for fusion algorithms. Some per-
formance metrics which are widely used in signal and image processing do not fit the
application of image fusion. One such example is mean square error. Let $f_1, f_2, ..., f_n :$
$[1, N] \times [1, M] \to \mathbb{R}$ be digital images of the same scene taken from different sensors and
f be the fusion result. A cost function characterizing the mean square error between
the fused image and source inputs may be defined as

$$\rho = \frac{1}{n} \sum_{i=1}^{n} \|f - f_i\|^2 \tag{6.18}$$

where $\| \cdot \|$ denote l_2 norm. It is easy to find out that fusion by pixel averaging, i.e.

$$f_a = \frac{1}{n} \sum_{i=1}^{n} f_i, \tag{6.19}$$

minimizes this cost. However, a pixel averaging method is not accepted as the best
fusion scheme [23]. It is noted earlier that pixel averaging has a muddy appearance.
This is due primarily to the fact that averaging results in reduced contrast for all the
patterns which appear in only one source and a mixture of different patterns which
come from different sources.

 Another possible candidate of performance measure for image fusion is the correlation
between the fused image and all the source inputs. The higher correlation, the better

	$C(f_1, f_2)$	$C(f_1, f_a)$	$C(f_2, f_a)$	$C(f_1, f_m)$	$C(f_2, f_m)$	$C(f_1, f_o)$	$C(f_2, f_o)$
		$C(f_1, f_a) + C(f_2, f_a)$		$C(f_1, f_m) + C(f_2, f_m)$		$C(f_1, f_o) + C(f_2, f_o)$	
Multi-modality	0.0498	0.5752	0.8457	0.4281	0.8703	0.4867	0.8405
		1.4208		1.2985		1.3272	
Multi-sensor	-0.0739	0.5790	0.7703	0.7928	0.3991	0.7663	0.4455
		1.3493		1.1919		1.2118	
Multi-spectral	0.8089	0.9501	0.9519	0.8959	0.9345	0.9231	0.9332
		1.9020		1.8303		1.8563	
Multi-focus	0.9542	0.9884	0.9886	0.9791	0.9766	0.9871	0.9778
		1.9770		1.9557		1.9649	

Table 6.1: Correlations between different fusion results and source images. f_1, f_2 denote two source images. f_a, f_m, f_o denote fusion outputs by pixel averaging, wavelet-based maximum selection and the proposed method respectively.

performance. Let $f, g : [1, N] \times [1, M] \to \mathbb{R}$ denote two digital images. The correlation between f and g is defined as

$$C(f, g) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} (f(i, j) - \bar{f})(g(i, j) - \bar{g})}{\sqrt{\sum_{i=1}^{N} \sum_{j=1}^{M} (f(i, j) - \bar{f})^2 \sum_{i=1}^{N} \sum_{j=1}^{M} (g(i, j) - \bar{g})^2}} \qquad (6.20)$$

where $\bar{f}, \bar{g}$ stands for the mean value of f, g respectively. Then a performance metric based on correlation may be defined as

$$\rho = \sum_{i=1}^{n} |C(f, f_i)|. \qquad (6.21)$$

where f is the fusion output and f_i's are source images. However, maximizing correlation is closely related to minimizing the mean square error. Fusion by pixel averaging usually maximizes the overall correlation given by Equation (6.21). Table (6.1) lists the correlations between different fusion outputs and source images in the experiments of multi-sensor, multi-modality, multi-spectral and multi-focus image fusion. Following the same argument for the cost function of mean square error, we conclude that correlation is also not a good choice to measure the performance of image fusion.

If the "ground truth" of the fusion result is known, we can perform a quantitative performance measure to compare different fusion algorithms. For the above experiment of multi-focus fusion, an ideal image should contain both well focused clocks and it may be constructed manually by cut and paste, as demonstrated in Figure (6.6.5).

Let $f_d, f : [1, N] \times [1, M] \to \mathbb{R}$ denote the ideal image and fusion result respectively. A performance measure ρ can be defined as the standard deviation of the error between f_d and f,

$$\rho = \sqrt{\frac{\sum_{i=1}^{N} \sum_{j=1}^{M} |f_d(i,j) - f(i,j)|^2}{NM}}. \tag{6.22}$$

Table (6.2) summarizes the standard deviation between an ideal image and fusion results by different algorithms. The proposed information theoretic approach clearly generates a fusion result that is the closest to the ideal image among the outputs of the pixel averaging and wavelet-based maximum selection scheme.

It has to be pointed out that this method is restricted to specially constructed images and generally not applicable to real multi-sensor data where the ideal fusion is ill-defined and cannot be obtained. There is no general quantitative performance measure for image fusion algorithms in the current literature except for specific applications [96]. The fusion results are mostly evaluated visually.

	pixel averaging	maximum selection	proposed algorithm
ρ	7.816	10.794	**6.573**

Table 6.2: Standard deviation between an ideal image and fusion results by different algorithms.

6.5 Conclusion

In this chapter, we derive a new multiscale image fusion algorithm, which aims at integrating complementary information from multi-sensor data so that the fused images

are more suitable for visual perception. We formulate the image fusion as an optimization problem to which we propose a solution.

As a first step, a biorthogonal wavelet transform of each source image is calculated to generate a scale space representation. Biorthogonal wavelets can be synthesized with perfect reconstruction filters having a linear phase, which is a desirable property for image fusion applications.

In contrast to the "choose max" type of selection rules, our proposed technique relies on the intrinsic statistical structure. Using spatially specific wavelet coefficients from fine to coarse scales, we construct activity pattern vectors which we compare using a new Jensen-Rényi divergence. The resulting decision map makes our approach more effective in preserving significant features from all the sources without suffering from artifacts.

We have successfully tested the new scheme on fusion of multi-sensor (low-light-television and forward-looking-infrared), multi-modality (CT and MRI), multi-spectral, and multi-focus images. Quantitative performance measure for fusion of synthetic test images and visual evaluation for real multi-sensor image fusion demonstrates that the presented algorithm clearly outperforms pixel averaging and wavelet based maximum selection fusion schemes.

(1)

(2)

(3)

(4)

(5)

(6)
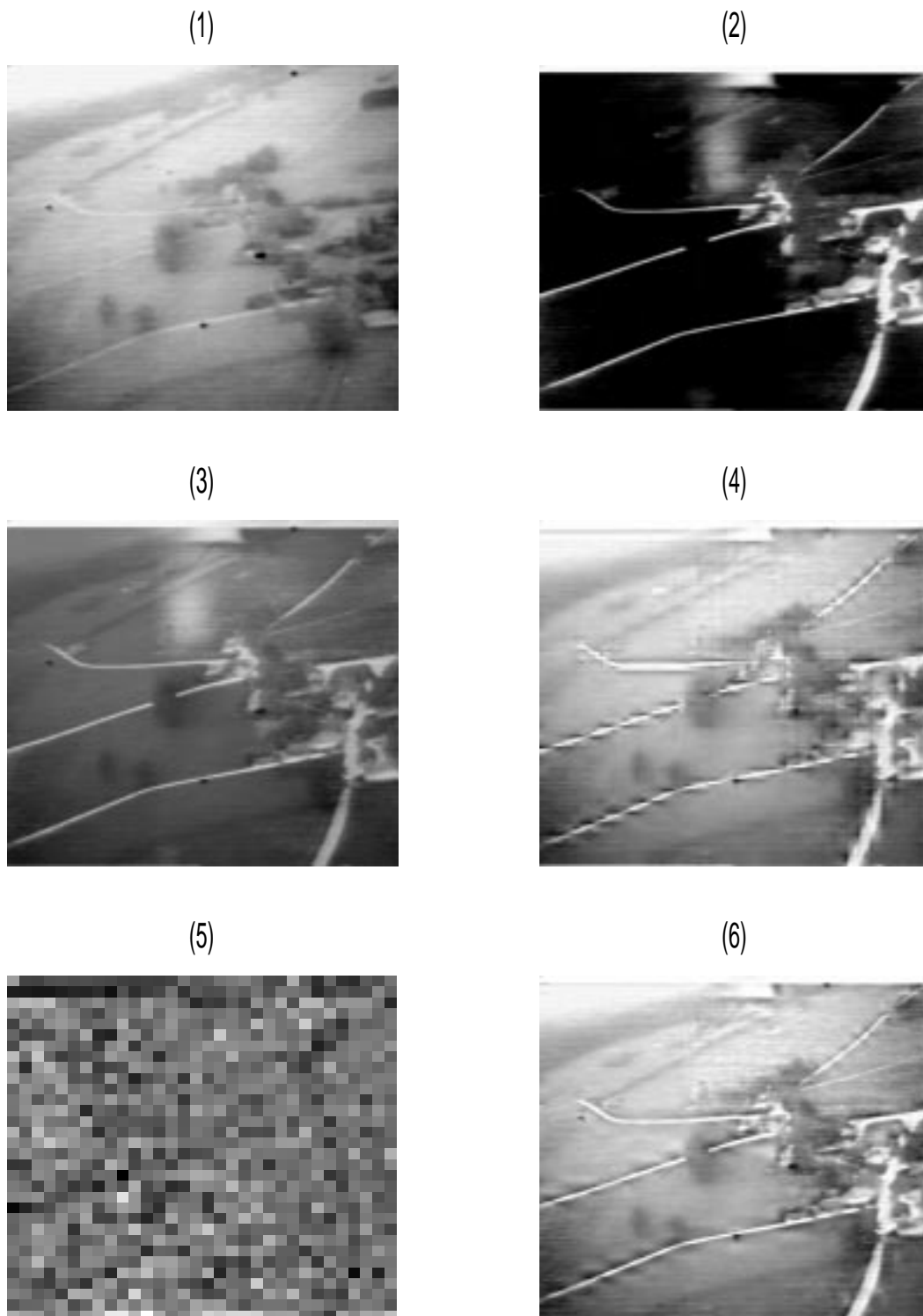
Figure 6.3: Multi-sensor image fusion: (1) a low-light-television sensor image; (2) a forward-looking-infrared image; (3) fusion by averaging; (4) fusion by wavelet based maximum selection scheme; (5) a selection map; (6) fusion by the proposed information theoretic approach.
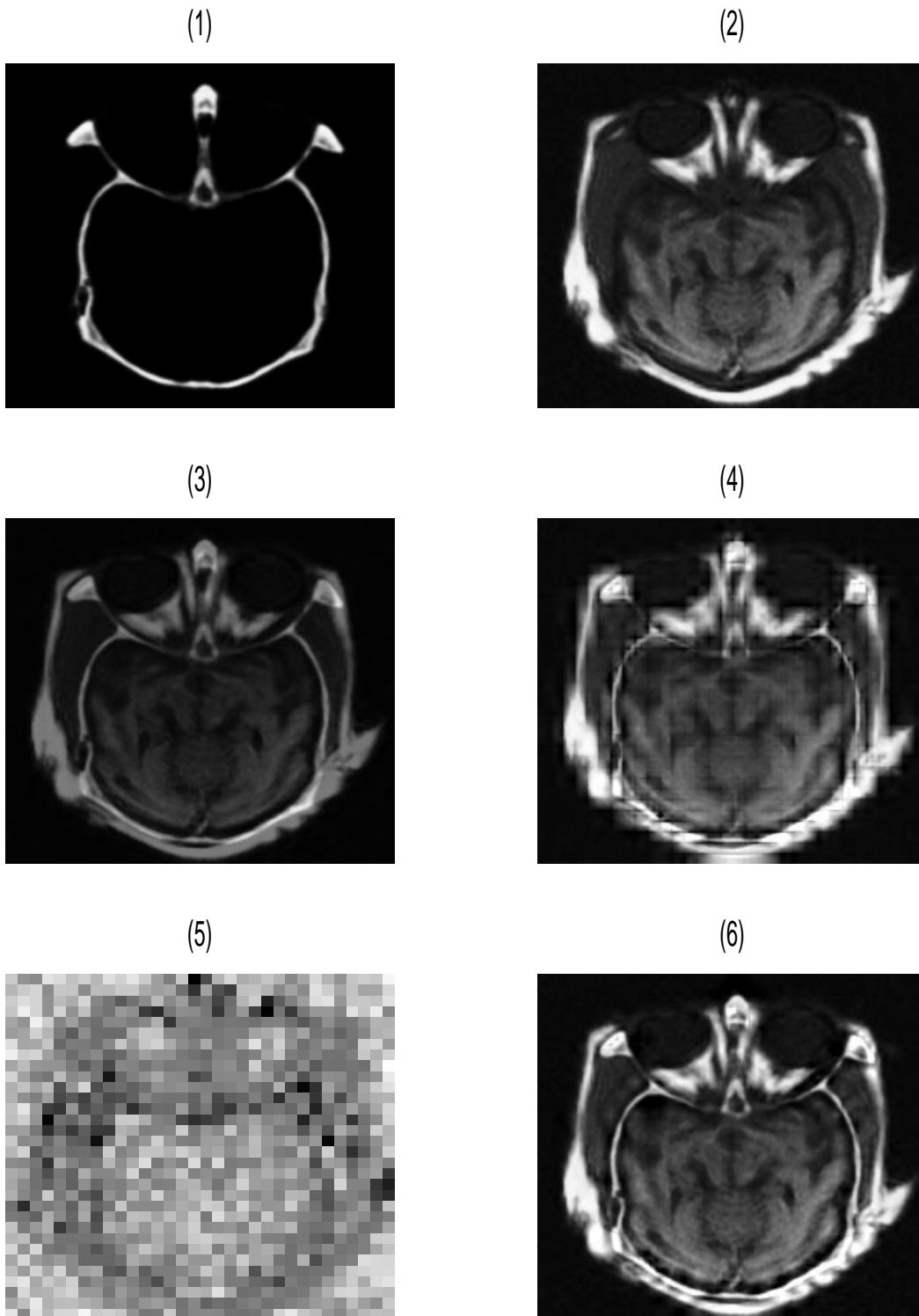
Figure 6.4: Multi-modality image fusion: (1) a CT image; (2) a MRI image; (3) fusion by averaging; (4) fusion by wavelet based maximum selection scheme; (5) a selection map; (6) fusion by the proposed information theoretic approach.

(1)   (2)

(3)   (4)

(5)   (6)
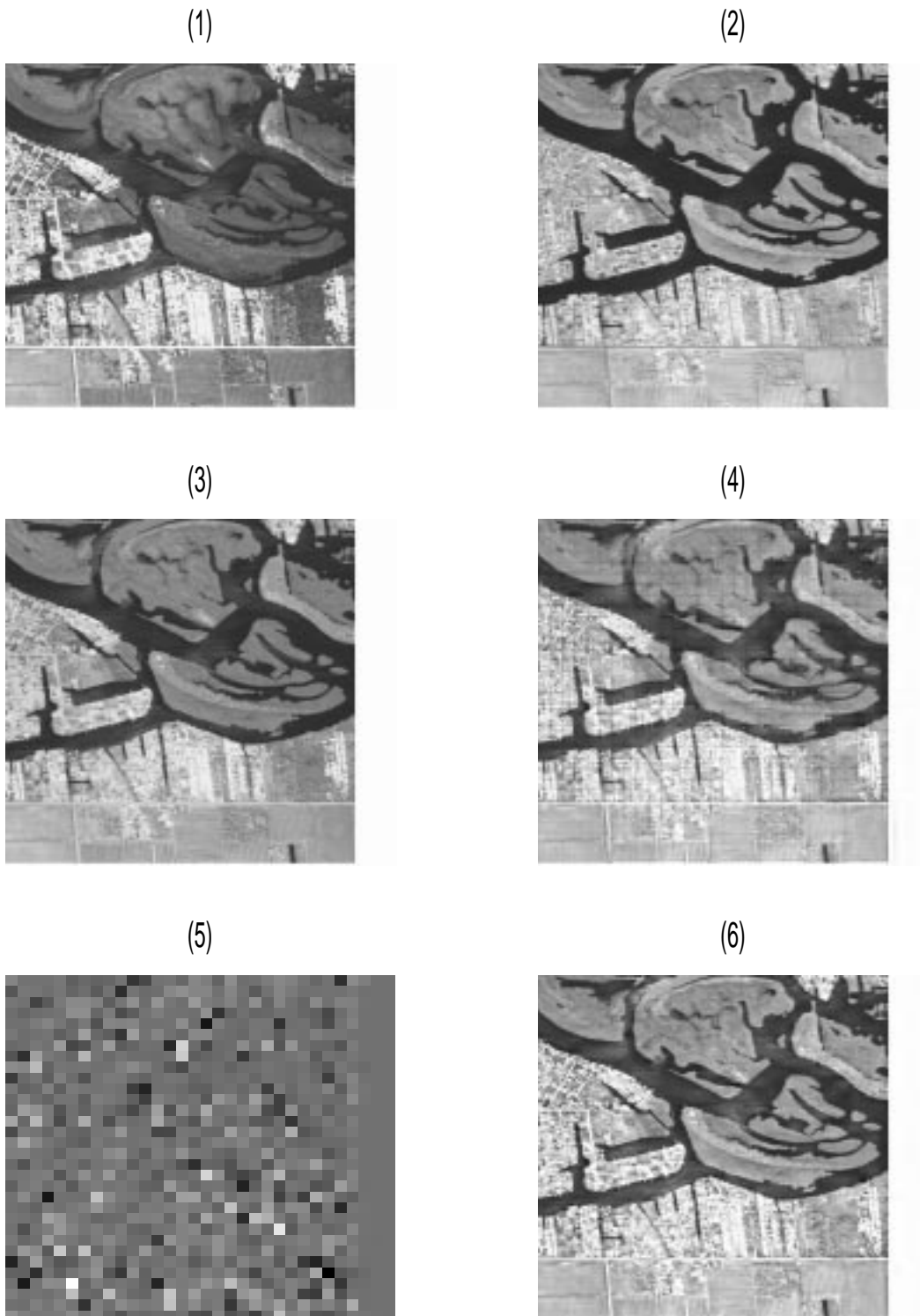
Figure 6.5: Multi-spectral image fusion: (1) a high resolution remote sensing image; (2) a low resolution remote sensing image; (3) fusion by averaging; (4) fusion by wavelet based maximum selection scheme; (5) a selection map; (6) fusion by the proposed information theoretic approach.
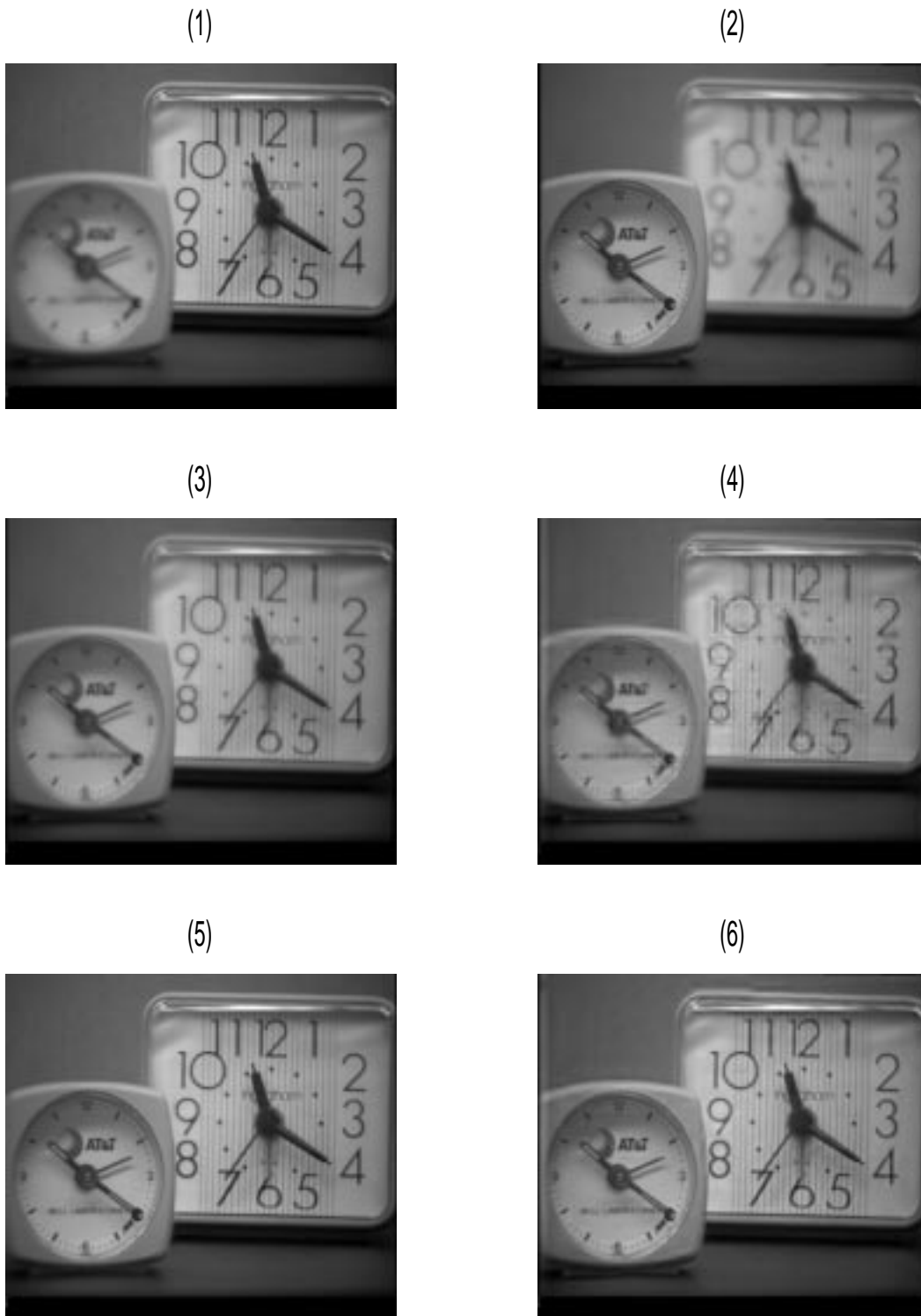
Figure 6.6: Multi-focus image fusion: (1) an image focused on the larger clock; (2) an image focused on the smaller clock; (3) fusion by averaging; (4) fusion by wavelet based maximum selection scheme; (5) a perfectly fused image obtained by manually cut and paste; (6) fusion by the proposed information theoretic approach.

7

Shape Recognition

THE geometrical description of an object can be decomposed into registration and shape information. For example, an object's location, rotation and size could be the registration information and the geometrical information that remains is the shape of the object. An object's shape is invariant under registration transformations and two objects have the same shape if they can be registered to match exactly.

The pioneers of this topic of general shape and registration analysis are Kendall [30] and Bookstein [31]. Some reference and reviews include Goodall [32], Kent [33], Dryden and Mardia [34].

The definition of general shape spaces and shape distance is introduced in Section 7.1. In Section 7.2 we describe the matching of two configurations under Euclidean similarity transformations. In Section 7.3 we briefly describe the matching of two configurations under affine transformations. We extend the work to consider the robustness issues and matching by estimation in Section 7.4.

7.1   Shape Space and Shape Distance

In this section, we investigate further geometrical aspects of shape. First we introduce the concepts of a shape configuration and of a pre-shape space.

Definition 7.1 *A configuration X consists of k labeled points in a m-dimensional Euclidean space.*

Two configurations, X and Y, have the same general shape if $Y = g(X)$, $g \in G$. G is the set of Euclidean transformations, including translation, rotation and isotropic scaling, which is given by

$$Y = rX\Gamma + \mathbf{1}_k c^T : r \in \mathbb{R}^+, \Gamma \in \mathbf{SO}(m), c \in \mathbb{R}^m \tag{7.1}$$

where c is a translation m-vector, $\mathbf{1}_k$ is k-vector of ones, Γ is a $m \times m$ orthogonal matrix with $\det(\Gamma) = 1$ and $r > 0$ denotes an isotropic scaling.

In order to represent shape, it is convenient to remove similarity transforms one at a time. Translation is the easiest to filter out of X and can be achieved following its product by a Helmert sub-matrix.

$$X_H = HX \in \mathbb{R}^{(k-1)m} \backslash \{0\} \tag{7.2}$$

A Helmert sub-matrix H is a $(k-1) \times k$ Helmert matrix without its first row. A full Helmert matrix H^F is a square $k \times k$ orthogonal matrix with its first row element equal to $1/\sqrt{k}$, and the remaining rows are orthogonal to the first row. We drop the first row of H^F, so that the transformed HX does not depend on the original locations of the configuration.

Definition 7.2 *The jth row of the Helmert sub-matrix H is given by*

$$(h_j, h_j, ...h_j, -jh_j, 0, ...0), \quad h_j = -(j(j+1))^{-1/2} \tag{7.3}$$

so that the jth row consists of h_j repeated j times, followed by $-jh_j$ and then $k - j - 1$ zeros, $j = 1, 2, ...k - 1$.

For example, a full Helmert matrix for $k = 3$ is

$$
H^F = \begin{bmatrix}
1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\
-1/\sqrt{2} & 1/\sqrt{2} & 0 \\
-1/\sqrt{6} & -1/\sqrt{6} & 2/\sqrt{6}
\end{bmatrix},
$$

while its corresponding Helmert sub-matrix is

$$
H = \begin{bmatrix}
-1/\sqrt{2} & 1/\sqrt{2} & 0 \\
-1/\sqrt{6} & -1/\sqrt{6} & 2/\sqrt{6}
\end{bmatrix}.
$$

The pre-shape of a configuration matrix X has all the information about location and scale removed.

Definition 7.3 *The pre-shape of a configuration matrix X is given by*

$$
Z = \frac{X_H}{\| X_H \|} = \frac{HX}{\| HX \|}
\tag{7.4}
$$

where H is a Helmert sub-matrix. The pre-shape is invariant under translation and scaling of the original configuration.

Notice that a pre-shape of configuration X has a dimension of $(k - 1)m - 1$. In order to remove rotation information from the configuration, we identify all rotated versions of the pre-shape with each other, and this equivalence class is the shape of X. An alternative definition of the shape of X is given by the following definition.

Definition 7.4 *The shape of a configuration matrix X is all the geometrical information about X that is invariant under Euclidean similarity transformations, i.e., location, rotation and isotropic scaling. The shape can be represented by the set $[X]$ given by*

$$
[X] = \{Z\Gamma : \Gamma \in \mathbf{SO}(m)\}
\tag{7.5}
$$

where $\mathbf{SO}(m)$ is the orthogonal group of rotations and Z is the pre-shape of X.

Definition 7.5 *The shape space [97] is the set of all the possible shapes. Formally, the shape space $\sum_m^k$ is the orbit space of the non-coincident k-point set configurations in $\mathbb{R}^m$ under the action of the Euclidean similarity transformations.*

A suitable distance in $\sum_k^m$ is the full Procrustean distance. Consider two configuration matrices X_1 and X_2 with pre-shapes Z_1 and Z_2, we minimize over rotations and scale to find the closest Euclidean distance between Z_1 and Z_2.

Definition 7.6 *The full Procrustean distance [34] between X_1 and X_2 is given by*

$$d_F(X_1, X_2) = \inf_{\Gamma \in \mathbf{SO}(m), \beta \in \mathbb{R}} \|Z_2 - \beta Z_1 \Gamma\| \tag{7.6}$$

It can be shown [30] that

$$\begin{aligned} d_F(X_1, X_2) &= \inf_{\Gamma \in \mathbf{SO}(m), \beta \in \mathbb{R}} \|Z_2 - \beta Z_1 \Gamma\| \\ &= \sqrt{1 - [\text{trace}(\Lambda)]^2} \\ &= \sin\rho. \end{aligned} \tag{7.7}$$

where ρ is a Riemannian metric, which is given by

$$\rho(Z_1, Z_2) = \arccos(\text{trace}(\Lambda)), \tag{7.8}$$

where the matrix Λ is the diagonal $m \times m$ matrix with the positive elements given by the square root of the eigenvalues of $Z_1^T Z_2 Z_2^T Z_1$, except the smallest diagonal element which is negative if we have $\det(Z_2^T Z_1) < 0$.

Another general shape space of interest is the affine shape space. The set of affine transformations of a configuration X ($k \times m$ matrix) is given by

$$\{XA + \mathbf{1}_k c^T : A \in \mathbf{GL}(m)\},$$

where A is a $m \times m$ matrix in the general linear group of invertible $m \times m$ matrices $\mathbf{GL}(m)$, and c is a translation m-vector. Note if $k \leq m + 1$ all the configurations have the same affine shape, so we require $k > m + 1$ for non-trivial affine shape.

Definition 7.7 *The affine shape space is the orbit space of the non-coincident k-point set configurations in $\mathbb{R}^m$ under the action of Euclidean affine transformations.*

A suitable distance in the affine shape space between X_1 and X_2 is given by

$$d_A(X_1, X_2) = \inf_{A \in \mathbf{GL}(m), c \in \mathbb{R}} \|X_1(X_1^T X_1)^{-1/2} - (X_2 A + \mathbf{1}_k c^T)\| \tag{7.9}$$

The post-multiplication of X_1 by $(X_1^T X_1)^{-1/2}$ ensures that $d_A(X_1, X_2) = d_A(X_2, X_1)$.

7.2 Euclidean Shape Matching

Consider configuration of $k \geq m + 1$ points in m dimension. To match two configurations Y and T (both $k \times m$ matrices), we have

$$Y = rT\Gamma + \mathbf{1}_k c^T + E = XB + E \tag{7.10}$$

where $c \in \mathbb{R}^m$ is a translation m-vector. $\mathbf{1}_k$ is the k-vector ones, Γ is a $m \times m$ orthogonal matrix with $\det(\Gamma) = 1$ and $r > 0$ denotes an isotropic scaling. E is the $k \times m$ error matrix, which is assumed to be modeled by an independent multivariate normal distribution. Let $X = [\mathbf{1}_k, T]$ be the design matrix, and $B = [c, r\Gamma]^T$ be the regression parameters. The most straightforward approach to estimating the regression parameters B is by a least squares approach, i.e., minimizing $s^2(E) = \|E\|^2 = \text{trace}(E^T E)$. To simplify calculation, we first obtain the pre-shapes of Y and T, denoted by $\check{Y}$ and $\check{T}$ respectively. In this way, c is set to be 0. Then we have

$$\begin{aligned} d_{GF}^2(Y, T) &= d_{GF}^2(\check{Y}, \check{T}) \\ &= \inf_{\Gamma \in \mathbf{SO}(m), r} \left((\check{Y} - r\check{T}\Gamma)^T (\check{Y} - r\check{T}\Gamma) \right) \\ &= \inf_r \left(\|\check{Y}\|^2 + r^2 \|\check{T}\|^2 - 2r \sup_{\Gamma \in \mathbf{SO}(m)} \text{trace}(\check{Y}^T \check{T}\Gamma) \right) \end{aligned} \tag{7.11}$$

Consider a singular value decomposition of $\check{Y}^T \check{T}$ given by

$$\check{Y}^T \check{T} = V \Lambda U^T$$
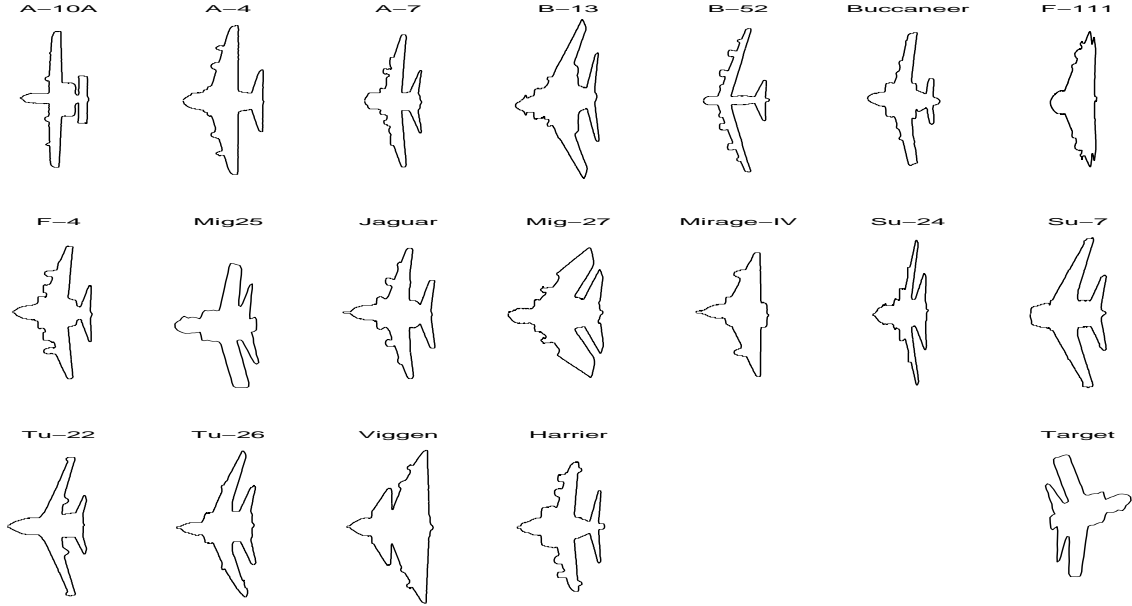
Figure 7.1: A shape database in experiment

where $U, V \in \mathbf{SO}(m)$, $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_m)$ with $\lambda_1 \geq \lambda_2, ..., \geq \lambda_{m-1} \geq |\lambda_m|$. Note that $\lambda_m < 0$ if and only if $\det(\check{Y}^T \check{T}) < 0$. Hence the last term in the right-hand side of Equation (7.11) is equivalent to

$$-2r \sup_{\Gamma \in \mathbf{SO}(m)} \mathbf{trace}(\Gamma \Lambda) = -2r \sup_{\Gamma \in \mathbf{SO}(m)} \sum_{i=1}^{m} \Gamma_{ii} \lambda_i,$$

where $(\Gamma_{11}, \Gamma_{22}, ..., \Gamma_{mm})$ are the diagonals of Γ. Note that the set of diagonals of Γ in $\mathbf{SO}(m)$ is a compact convex set [98] with extreme points

$$\{(\pm 1, \pm 1, ..., \pm 1)\}$$

with an even number of minus signs. Hence it is clear in our case that the supremum is achieved when $\Gamma_{ii} = 1$, $i = 1, 2, ..., m$. Therefore,

$$d_{GF}^2(\check{Y}, \check{T}) = \inf_r \left(\|\check{Y}\|^2 + r^2 \|\check{T}\|^2 - 2r \sum_{i=1}^{m} \lambda_i \right). \tag{7.12}$$

By differentiation we obtain

$$\hat{r} = \sum_{i=1}^{m} \lambda_i. \tag{7.13}$$

It is now clear the minimizing rotation matrix is given by $\hat{\Gamma} = UV^T$ since

$$\text{trace}(\check{Y}^T \check{T} \hat{\Gamma}) = \text{trace}(V \Lambda U^T U V^T) = \text{trace}(\Lambda).$$

Based on the above derivation, a suitable symmetric residual discrepancy measure involves the matching of the pre-shapes under Euclidean similarity transformations and is given by

$$d_{GF}(Y, T) = \inf \| \check{Y} - [\mathbf{1}_k, \check{T}][c, r\Gamma]^T \| = \sqrt{(1 - (\text{trace}(\Lambda))^2)}. \tag{7.14}$$

Referring to equation (7.7), we can see that the symmetric residual discrepancy measure corresponds to a full Procrustean distance in shape space.

An aircraft database of 18 templates and a target configuration we extracted from an ISAR image are shown in Figure (7.1). The target configuration can be modeled by a translation of MIG-25 in the database plus some distortions. We center and normalize it to a pre-shape format. Symmetric discrepancy measures between the target and templates in database are calculated and listed in the Table (7.1). As we can see, the shape distance between the target and MIG-25 has the minimal value.

	A-4	A-7	A-10	B-13	B-52	Buccaneer
$d_{GF}(Y, T)$	0.3416	0.3460	0.4339	0.2776	0.5492	0.2645
	F-111	F-4	MIG-25	Jaguar	MIG-27	Mirage-IV
$d_{GF}(Y, T)$	0.3653	0.3336	**0.0501**	0.2930	0.2737	0.3837
	Su-24	Su-7	Tu-22	Tu-26	Viggen	Harrier
$d_{GF}(Y, T)$	0.2425	0.3197	0.2491	0.2408	0.4086	0.3086

Table 7.1: Symmetric discrepancy measures between the target and aircraft templates.

7.3 Affine Shape Matching

For $k \geq m + 1$ points in $\mathbb{R}^m$, the set of transformations G is the group of affine transformations. Matching two configurations Y and T (both $k \times m$ matrices), we have a

$(m + 1) \times m$ affine transformation matrix B such that

$$Y = [\mathbf{1}_k, T]B + E = XB + E \tag{7.15}$$

where E is the $k \times m$ error matrix, which is assumed to be modeled by an independent multivariate normal distribution. $X = [\mathbf{1}_k, T]$ is the design matrix. The least square solution is given by

$$\hat{B} = (X^T X)^{-1} X^T Y, \tag{7.16}$$

where we assume the rank of X to be $m + 1$. A symmetric discrepancy measure can be obtained by

$$d_{GA}(Y, T) = \|Y(Y^T Y)^{-1/2} - X(X^T X)^{-1} X^T Y\|. \tag{7.17}$$

This measure has the property that

$$d_{GA}(Y, T) = d_{GA}(Y, TA + \mathbf{1}_k b^T) = d_{GA}(T, Y) = d_{GA}(YA + \mathbf{1}_k b^T, T)$$

assuming T, Y are of rank m, A is a full rank $m \times m$ matrix and b is a m-vector.

7.4   Estimation by Matching

If we have N random samples of a object, $T_j \in \sum_k^m, j = 1, 2, ...N$, then it is of interest to obtain an average configuration μ (a $k \times m$ matrix) and to explore the structure of viability, up to invariance in the set of transformations G.

An estimate of the population mean configuration μ up to invariance in G, denoted by $\hat{\mu}$, can be obtained by simultaneously matching each T_j to μ ($j = 1, 2, ...N$) and choosing μ as a suitable estimate (subject to certain constraints on μ). In particular, $\hat{\mu}$ is obtained from the constrained minimization

$$\hat{\mu} = \arg \inf_{\mu} \sum_{j=1}^{N} \inf_{g_j} \phi(s(g_j(T_j) - \mu)) \tag{7.18}$$

where $\phi(x)$ is a penalty function on $\mathbb{R}^+$, $s(E)$ is the objective function for matching two configurations, and in general restrictions need to be imposed on μ to avoid degeneracy. A common choice of estimator is $\phi(x) = x^2$, i.e., a least square choice.

For the generalized shape matching in two dimensions, given N random sample configurations in the complex plane $T_1, ..., T_N, (T_j \in \mathbb{C}^k$, complex k-vectors), we wish to obtain an average shape, where the set of transformations G is taken to be the Euclidean similarity group. Upon selecting $\phi(x) = x^2$ for a least square matching, we let $X_j = [\mathbf{1}_k, T_j], j = 1, 2, ..., N$ be a $k \times 2$ design matrix, and define

$$\epsilon_j = \mu - X_j B_j,$$

where B_j are the 2×1 complex parameters for matching the jth configuration to a complex mean μ (a $k \times 1$ complex vector) with μ restricted to be centered ($\mu^* \mathbf{1}_k = 0$) and having unit size ($\mu^* \mu = 1$). " * " denotes complex conjugate. ϵ_j is assumed to be modeled by an independent multivariate normal distribution.

The N configurations can hence be matched using a least square criterion, i.e., minimizing the function $\sum_{j=1}^N \epsilon_j \epsilon_j^*$ over $\{B_j\}_{j=1}^N$ and μ. Fixing μ, the solution is clearly given by

$$\hat{B}_j = (X_j^* X_j)^{-1} X_j^* \mu, \quad (j = 1, 2, ...N) \tag{7.19}$$

where $X_j^* X_j$ is assumed of rank 2, i.e., the points in T_j are not all coincident. This yields,

$$\hat{\epsilon} = (I_k - H_j)\mu, \tag{7.20}$$

where H_j is the estimated hat matrix for X_j and given by

$$H_j = X_j(X_j^* X_j)^{-1} X_j^*, \quad (j = 1, 2, ...N).$$

Consequently, we have [34],

$$\hat{\mu} = \arg\inf_{\mu} \mu^* A \mu, \tag{7.21}$$

where $A = \sum_{j=1}^N (I - H_j)$. Hence, $\hat{\mu}$ is given by a complex eigenvector corresponding to the smallest non-zero eigenvalue of A, which is equivalent to finding an eigenvector corresponding to the largest eigenvalue of
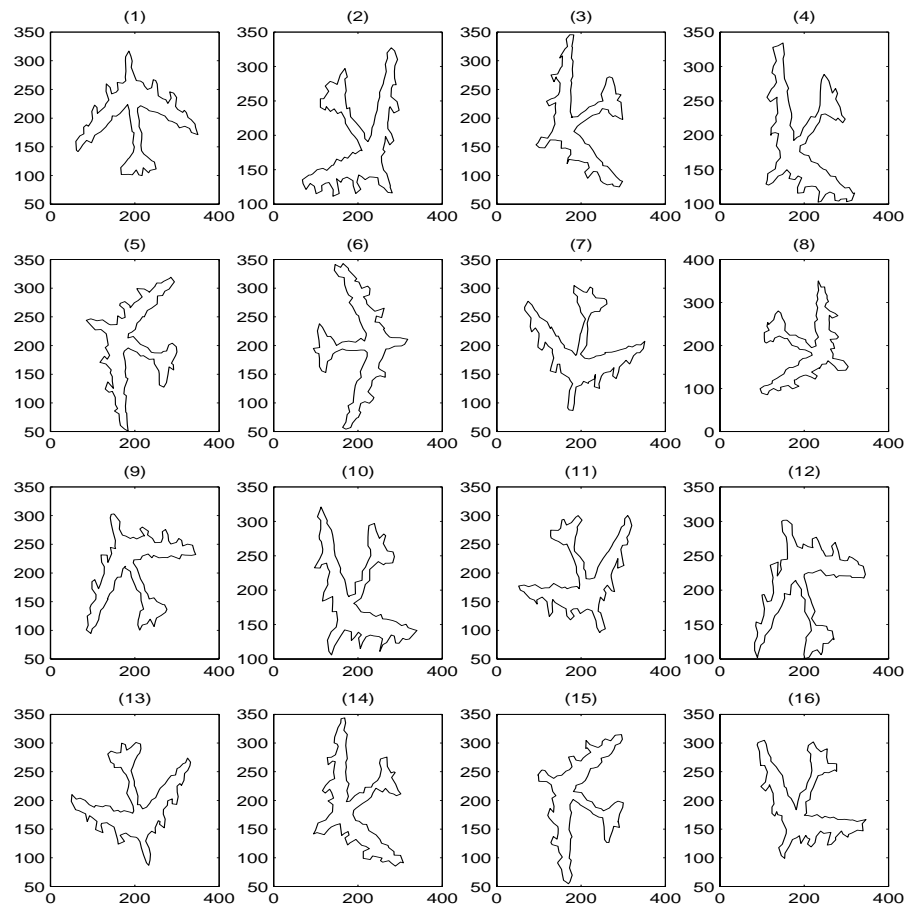
$$S = \sum_{j=1}^N \check{T}j\check{T}_j^*, \tag{7.22}$$
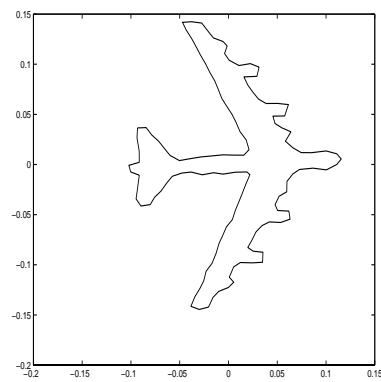
Figure 7.2: A set of 16 noisy observation of a B-52 shape

Figure 7.3: Estimated mean configuration of a B-52 shape

	A-4	A-7	A-10	B-13	B-52	Buccaneer
$d_{GF}(Y,T)$	0.4792	0.4091	0.3661	0.5675	**0.0081**	0.4249
	F-111	F-4	Harrier	Jaguar	MIG-27	Mirage-IV
$d_{GF}(Y,T)$	0.7045	0.5488	0.5173	0.6030	0.5850	0.7195
	Su-24	Su-7	Tu-22	Tu-26	Viggen	MIG-25
$d_{GF}(Y,T)$	0.4754	0.4936	0.4604	0.5477	0.6974	0.5494

Table 7.2: Shape distances between the estimated mean shape and templates in the database

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8
$d_{GF}(Y_i,T)$	0.0369	0.0417	0.0393	0.0376	0.0426	0.0399	0.0396	0.0397
	Y_9	Y_{10}	Y_{11}	Y_{12}	Y_{13}	Y_{14}	Y_{15}	Y_{16}
$d_{GF}(Y_i,T)$	0.0389	0.0385	0.0389	0.0408	0.0398	0.0398	0.0392	0.0421

Table 7.3: Shape distances between noisy configurations to the B-52 template.

where $\check{T}_j$ is the pre-shape of T_j. The shape corresponding to this eigenvector is hence the unique least square estimate provided that there is a single greatest eigenvalue. The average shape from this procedure is called the full procrustean mean.

Figure (7.2) shows a set of noisy configurations (128×1 complex vectors) of B-52. The estimated mean shape calculated by Equation (7.21) is shown in Figure (7.3). Table (7.2) lists the shape distances between the estimated mean shape and templates in the aircraft database. Table (7.3) lists the shape distances between noisy configurations to the B-52 template. It's clear that matching by way of a mean shape has a better result.

8 Conclusions and Discussions

IN this chapter, we briefly summarize the contributions of this thesis and the overall conclusions which can be derived from the results of our research. We also present some suggestions for extending this work.

8.1 Conclusions

The main contributions of this thesis can be categorized as follows:

• **Information theoretic approach for ISAR image focusing and multiscale data fusion**

A new generalized divergence measure, the *Jensen-Rényi* divergence, is proposed. We prove the convexity of this divergence measure, derive its maximum value, and analyze its upper bounds in terms of the Bayes error in statistical pattern recognition. Based on the Jensen-Rényi divergence, we propose a new approach to the problem of

ISAR image registration. This is accomplished by using the Jensen-Rényi divergence to measure the statistical dependence between consecutive ISAR image frames, which is maximal if the images are geometrically aligned. Compared to the mutual information based registration techniques, the Jensen-Rényi divergence provides an ability to control the measurement sensitivity of the joint histogram. This flexibility results in more robust registration in the presence of noise. Maximization of the Jensen-Rényi divergence is a very general criterion, because no assumptions are made in regards to the nature of this dependence and no limiting constraints are imposed on image contents. Simulation results demonstrate that our approach achieves an effective estimation of the target motion automatically without any prior feature extraction. Based on the estimated motion parameters, translational motion compensation (TMC), and rotational motion compensation (RMC) can be used to generate a focused image of the target.

In order to integrate complementary information from multi-sensor data so that the fused images are more suitable for visual perception and recognition, we derive a new information theoretic fusion scheme in a multiscale framework. We formulate the image fusion as an optimization problem for which we propose a solution. We have successfully tested the new scheme on fusion of multi-sensor (low-light-television and forward-looking-infrared), multi-modality (CT and MRI), multi-spectral, and multi-focus images. Quantitative performance measure for fusion of synthetic test images and visual evaluation for real multi-sensor image fusion demonstrates that the presented algorithm clearly outperforms pixel averaging and wavelet based maximum selection fusion schemes.

- **Multiscale shape enhancement and shape analysis**

As a pre-processing step for the shapes extracted from ISAR images, we propose a novel non-linear smoothness-constrained filtering technique. The key idea is to separate the signal portion from its measured data, and to preserve the original smoothness property of the underlying shape. Using notations of Hölder spaces and Hölder exponent, we establish results of signal regularity measurement with wavelets. To detect

the singular points of signal from measured data, we turned to curve shortening and derived the partial differential equations that characterize the evolution of curvature. A new singularity detection method by tracking the curvature extrema across scales is proposed and a multiscale curvature mask is generated. Then we proceed to project measured data into the wavelet domain and suppress wavelet coefficients by this multiscale curvature mask. For a piecewise smooth signal, it was shown that filtering by this curvature mask is equivalent to keeping the signal pointwise Hölder exponents at the singular points of the underlying signal, and to lifting its smoothness at all the remaining points.

To identify a shape independently of its registration information, we finally propose matching two configurations by regression, using notations of general shape spaces and procrustean distances. In particular, we study the generalized Euclidean and affine matching by estimating a mean configuration in two dimensions. Simulation results show that matching by way of a mean configuration is more robust than matching target shapes directly.

8.2   Suggestions for future research

In this section, we provide some discussions and suggestions for extending the research presented in this thesis.

• **Optimal choice of exponential order for the Jensen-Rényi divergence**

A large fraction of our effort in Chapter 3 was focused on searching for a spatial transformation such that a similarity metric achieves its maximum between two images taken at different times, from different sensors, or from different viewpoints. We propose a general framework based on the Jensen-Rényi divergence for the purpose of image registration and establish the optimal choice of weight vector ω. For the selection of exponential order α, there is a trade off between optimality and practicality. We conclude that $\alpha = 0$ is best for the ideal case of registration, if one can exactly model

the misalignment between two images. It is, however, also the least robust selection, as it tends to make all the $p_i's$ the same as the uniform distribution, if p_i is not degenerate distribution and $p_{ij} > 0$. Then the Jensen-Rényi divergence would be zero for the whole transformation parameter space as in case where the adapted transformation group can not completely model the misalignment. On the other hand, $\alpha = 1$ is the most robust choice, in spite of also resulting in the least sharp peak. In the future work, I would suggest formulating a cost function over α and solving its optimal value by minimizing that cost function. The difficulty is to find such a cost function that it is convex in α and also reasonable for a specific application.

• **Registration in the presence of local variation**

For the purpose of ISAR image focusing, local variations may be presented in the consecutive image frames due to the target motion and other perturbations. If such local variations are severe, a global transform and a local deformation may be used together to aid the Jensen-Rényi divergence to identify the registration point. Even though in the motion compensation stage, the local transformation is not required. The local deformation would also help to determine the maximal integration angle, since it indicates the projection variation of target reflectivity density onto the imaging plane. Given two ISAR image frames f_1 and f_2, an estimate of target motion parameters $(l^*, \theta^*, \gamma^*)$ is then given by

$$(l^*, \theta^*, \gamma^*) = \arg \max_{(l,\theta,\gamma)} D_{JR}{}^{\boldsymbol{\omega}}_{\alpha} \left(f_1, g(\mathcal{T}_{(l,\theta,\gamma)} f_2) \right)$$

where $D_{JR}{}^{\boldsymbol{\omega}}_{\alpha}(\cdot)$ is an induced similarity measure based on Jensen-Rényi divergence of order α and weight $\boldsymbol{\omega}$. g is a local deformation. A good candidate for such a local deformation is a thin-plate spline [34]. The thin-plate spline is a natural interpolating function for data in two dimensions and plays a similar role to the natural cubic spline in the one dimensional case. It minimizes the total bending energy of all possible interpolation functions [34]. Implementation of a thin-plate spline interpolation is straight forward. The difficulty could be a robust scheme to select two sets of feature landmarks to characterize the target in both images. These two sets of landmarks are necessary to

calculate the parameters in a thin-plate spline interpolation.

- **Segmentation techniques to construct decision maps in image fusion**

Improving the segmentation techniques to construct a decision map is crucial for multiscale image fusion. I would suggest applying region growing methods to segment the selection map in the future research. Region growing methods are well suited for the selection maps, where borders are difficult to detect. Computation cost used to be a problem, however, a recently developed fast watershed algorithm by Vincent and Soille [99] was found to be several hundreds of times faster than several classical algorithms. Image segmented by region growing methods sometimes contain either too many regions (under-growing) or too few regions (over-growing) as a result of non-optimal parameter setting. To improve classification results, a variety of post-processors [100] has been developed. I would also suggest investigating these post-processors to fine tune the resulting decision map. A well constructed decision map could reduce the artifacts introduced by switching source wavelet coefficients and hence improve the quality of the fusion result.

Bibliography

[1] Woods R. P.; Mazziotta J. C.; Cherry S. R., "Mri-pet registration with automated algorithm," *J. Comput. Assist. Tomogr.*, vol. 17, no. 4, pp. 536–546, 1993.

[2] P. Viola and W. M. Wells, "Alignment by maximization of mutual information," *International Journal of Computer Vision*, vol. 24, no. 2, pp. 137–154, 1997.

[3] F. Maes; A. Collignon; D. Vandermeulen; G. Marchal; P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Trans. on Medical Imaging*, vol. 16, no. 2, pp. 187–198, 1997.

[4] Hero A.O. III.; Michel O.J.J., "Asymptotic theory of greedy approximations to minimal k-point random graphs," *IEEE Trans. Information Theory*, vol. 45, no. 6, pp. 1921–1938, 1999.

[5] A. Rényi, "On measures of entropy and information," *Selected Papers of Alfréd Rényi*, vol. 2, pp. 525–580, 1976.

[6] Donald R. Wehner, *High Resolution Radar, 2nd edition*, Artech House, Artech House Inc., Norwood, MA 02062, USA, 1995.

[7] D. Donoho and I. Johnstone, "Ideal denoising in an orthogonal basis chosen from a library of bases," *C. R. Acad. Sci. Paris*, vol. 319, pp. 1317–1322, Oct. 1994.

[8] H. Krim and J-C Pesquet, *On the Statistics of Best Bases Criteria*, vol. Wavelets in Statistics of *Lecture Notes in Statistics*, Springer-Verlag, July 1995.

[9] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[10] D. Donoho and I. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1200–1224, 1995.

[11] M. Neumann and R. von Sachs and, *Wavelet thresholding:Beyond the gaussian iid situation*, vol. 103 of *Wavelets and Statistics, Lecture Notes in Statistics*, Springer-Verlag, Nov. 1994.

[12] R. Averkamp and C. Houdre, "Wavelet thresholding for non gaussian noise: idealism," *Preprint*, pp. 1–64, 1999.

[13] H. Krim and I. C. Schick, "Minimax description length for signal denoising and optimized representation," *IEEE Trans. Information Theory*, vol. 45, no. 3, pp. 898–908, 1999.

[14] Peter R.Massopust, *Fractal Functions, Fractal Surfaces, and Wavelets*, Academic Press, Academic Press Limited, London, UK, 1994.

[15] S. Jaffard, "Pointwise smoothness, two-microlocalization and wavelet coefficients," *Publications Mathématiques*, vol. 35, pp. 155–168, 1991.

[16] Y. Meyer, *Wavelets and Applications*, Cambridge University Press, first edition, 1992.

[17] Ajjimarangsee P.; Huntsberger T. L., "Neural network model for fusion of visible and infrared sensor outputs," *Proc. SPIE*, vol. 1003, pp. 153–160, 1988.

[18] Lee J. S. J., "Multiple sensor fusion based on morphological processing," *Proc. SPIE*, vol. 1003, pp. 94–100, 1988.

[19] Ritter G.X.; Wilson J. N. and Davidson J. L., "Image algebra application to multisensor and multi-data image manipulation," *Proc. SPIE*, vol. 933, pp. 2–7, 1988.

[20] Chen S., "Stochastic image algebra for multi-sensor fusion and spatial reasoning: a neural approach," *Proc. SPIE*, vol. 1098, pp. 146–154, 1989.

[21] Huntsberger T. L., *Data fusion: A neural networks implementation*, Data fusion in robotics and machine intelligence, M. Abidi and R. Gonzalez Eds., Academic Press, San Diego, 1992.

[22] Wright W. A., "A markov random field approach to data fusion and color segmentation," *Image Vision Comp.*, vol. 7, no. 2, pp. 144–150, 1989.

[23] R. Luo and M. Kay, *Data fusion and sensor integration: state of the art in 1990s*, Data fusion in robotics and machine intelligence, M. Abidi and R. Gonzalez Eds., Academic Press, San Diego, 1992.

[24] P.J. Burt and R. J. Kolczynski, "Enhanced image capture through fusion," *Proc. 4th Intl Conference on Computer Vision*, pp. 173–182, 1993.

[25] Li H.; Manjunath B. S. and Mitra S. K., "Multisensor image fusion using the wavelet transform," *Graphical Models and Image Processing*, vol. 57, no. 3, pp. 235–245, 1995.

[26] Yocky D. A., "Artifacts in wavelet image merging," *Optical Eng.*, vol. 35, no. 7, pp. 2094–2101, 1996.

[27] Rockinger O., "Pixel level fusion of image sequences using wavelet frames," *Proc. 16th Leeds Annual Statistical Research Workshop*, pp. 149–154, 1996.

[28] P.J. Burt, *The Pyramid as a Structure for Efficient Computation*, Multiresolution image processing and analysis, A. Rosenfeld Ed., Springer-Verlag, New York, 1984.

[29] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Commun. on Pure and Appl. Math*, vol. 41, pp. 909–996, 1988.

[30] D.G. Kendall, "A survey of the statistical theory of shape," *Statistical Science*, vol. 4, pp. 87–120, 1989.

[31] F.L. Bookstein, "Thin plate splines and the decomposition of deformations," *IEEE Trans. Pattern Analysis, and Machine Intelligence*, vol. 11, no. 7, pp. 567–585, 1989.

[32] C.R. Goodall, "Procrustes methods in the statistical analysis of shape (with discussion)," *Journal of the Royal Statistical Society B*, vol. 53, pp. 285–339, 1991.

[33] J.T. Kent, "New directions in shape analysis," *The Art of Statistical Analysis*, pp. 115–127, 1992.

[34] Ian L. Dryden and Kanti V. Mardia, *Statistitcal Shape Analysis*, John Willey Sons, John Willey Ltd, Baffins Lane, UK, 1998.

[35] Yun He; A. Ben Hamza; H. Krim; V.C. Chen, "An information theoretic measure for isar imagery focusing," *Proc. SPIE*, vol. 4116, pp. 463–471, 2000.

[36] Yun He; A. Ben Hamza; H. Krim, "Image registration for interferometric inverse sar focusing," *IEEE Statistical Signal Proc. Workshop*, 2001.

[37] Yun He; A. Ben Hamza; Hamid Krim, "An information divergence measure with its application to image registration," *submitted to IEEE Trans. on Signal Processing*, 2001.

[38] Yun He and Hamid Krim, "Multiscale shape analysis: Beyond the normality and independence of noise," *Proc. SPIE*, vol. 3813, pp. 233–244, 1999.

[39] Yun He and Hamid Krim, "Multiscale signal enhancement: Beyond the normality and independence assumption," *submitted to IEEE Trans. on Image Processing*, 2001.

[40] Yun He; A. Ben Hamza; Hamid Krim, "A multiscale approach for pixel level image fusion," *in preparation*, 2001.

[41] David C. Munson and Robert L. Visentin, "A signal processing view of strip-mapping synthetic aperture radar," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 37, no. 12, pp. 2131–2147, 1989.

[42] Katuri R. and Jain R. C., *Computer Vision: Principles*, IEEE Computer Society Press, Los Alamitos, CA, 1991.

[43] Horn B. K. P., *Robot Vision*, MIT Press, Cambridge, MA, 1989.

[44] Jensen J. R., *Introductory digital image processing : a remote sensing perspective 2nd edition*, Prentice Hall, Upper Saddle River, NJ, 1996.

[45] Thomas I.L.; Benning V. M.; Ching N. P., *Classification of remotely sensed images*, Adam Hilger, Bristol, England, 1986.

[46] Stytz M.R.; Frieder S.; Frieder O., "Three dimensional medical imaging: Algorithms and computer systems," *ACM Computing Surveys*, vol. 23, no. 4, pp. 421–424, 1991.

[47] Van den Elsen P.A.; Pol E.-J.D.; Viergever M.A., "Medical image matching-a review with classification," *IEEE Engineering in Medicine and Biology Magazine*, vol. 12, no. 1, pp. 26–39, 1993.

[48] Gabriel A. K.; Goldstein R. M., "Crossed obit interferometry: Theory and experimental results from sir-b," *Int. J. Remote Sens.*, vol. 9, pp. 857–872, 1988.

[49] Li F.K.; Goldstein R.M., "Studies of multibaseline spaceborne interferometric synthetic aperture radars," *IEEE Trans. Geoscience and Remote Sensing*, vol. 28, no. 1, pp. 88–97, 1990.

[50] Lin Q.; Vesecky J.F.; Zebker H.A., "New approaches in interferometric sar data processing," *IEEE Trans. Geoscience and Remote Sensing*, vol. 30, no. 3, pp. 560–567, 1992.

[51] G. Fornaro; G. Franceschetti, "Image registration in interferometric sar processing," *Radar, Sonar and Navigation, IEE Proceedings*, vol. 142, no. 6, pp. 313–320, 1995.

[52] L. Brown, "An survey of image registration techniques," *ACM Computing Surveys*, vol. 24, no. 4, pp. 325–376, 1992.

[53] Rosenfeld A.; Kak A.C., *Digital Picture Processing*, Academic Press, Orlando, Fla., 1982.

[54] Kuglin C. D.; Hines D. C., "The phase correlation image alignment method," *Proceedings of the IEEE 1975 international conference on cybernetics and society*, pp. 163–165, 1975.

[55] De Castro E.; Morandi C., "Registration of translated and rotated images using finite fourier transforms," *IEEE Trans. Pattern Analysis, and Machine Intelligence*, vol. 9, pp. 700–703, 1987.

[56] Goshtasby A., "Image registration by local approximation," *Image Vision Comput.*, vol. 6, no. 4, pp. 255–261, 1988.

[57] Broit C., "Optimal registration of deformed images," *Ph.D. Thesis, Univ. of Pennsylvania*, 1981.

[58] Bajscy R.; Kovacic S., "Multiresolution elastic matching," *Comput. Vission Graph. Image Processing*, vol. 46, pp. 1–21, 1989.

[59] Burr D. J., "A dynamic model for image registration," *Comput. Graphics Image Proc.*, vol. 15, pp. 102–112, 1981.

[60] C.E. Shannon, "A mathematical theory of communication," *Bell Systems Tech Journal*, vol. 27, pp. 379–423, 1948.

[61] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Trans. Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.

[62] A.W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, 1979.

[63] M. Hellman and J. Raviv, "Probability of error, equivocation, and the chernoff bound," *IEEE Trans. Information Theory*, vol. 16, no. 4, pp. 368–372, 1970.

[64] T.M. Cover and P.E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Information Theory*, vol. 13, pp. 21–27, 1967.

[65] Victor C. Chen and Shie Qian, "Joint time frequency transform for radar range-doppler imaging," *IEEE Trans. Aerospace and Electronic Systems*, vol. 34, no. 2, pp. 486–499, 1998.

[66] G. Gallager, *Information Theory and Reliable Communications*, John Willey Sons, John Willey Ltd, Baffins Lane, UK, 1968.

[67] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Analysis, and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[68] P.J. Burt; E.H. Adelson, "The laplacian pyramid as a compact image code," *IEEE Trans. Communications*, vol. 31, no. 4, pp. 532–540, 1983.

[69] J. G. Proakis; D. G. Manolakis, *Digital Signal Processing - Principles, Algorithms, and Applications, 3rd edition*, Prentice Hall, Upper Saddle River, NJ, 1996.

[70] Stephane Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, Academic Press Limited, London, UK, 1997.

[71] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, PA, 1992.

[72] O. Rioul, "Regular wavelets: A discrete-time approach," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3572–3578, 1993.

[73] A. Croisier; D. Esteban; C. Galand, "Perfect channel splitting by use of interpolation-decimation tree decomposition techniques," *Int. Conf. on Info. Sciences and Systems, Patras, Greece*, pp. 443–446, 1976.

[74] M.J. Smith; T. P. Barnwell III, "A procedure for designing exact reconstruction filter banks for tree structured sub-band coders," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1984.

[75] F. Mintzer, "Filters for distortion free two band multirate filter banks," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 33, no. 3, pp. 626–630, 1985.

[76] M. Vetterli, "Splitting a signal into subsampled channels allowing perfect reconstruction," *Proc. IASTED Conf. on Appl. Sig. Proc. and Dig. Filt., Paris, France*, 1985.

[77] M. Vetterli, "A filter banks allowing perfect reconstruction," *Signal Process.*, vol. 10, no. 3, pp. 219–244, 1986.

[78] A. Cohen; I. Daubechies; J. C. Feauveau, "Biorthogonal bases of compactly supported wavelets," *Commun. on Pure and Appl. Math.*, vol. 45, pp. 485–560, 1992.

[79] C. Herley; M. Vetterli, "Wavelets and recursive filter banks," *IEEE Trans. Signal Process.*, vol. 41, no. 8, pp. 2536–2556, 1993.

[80] W. Sweldens, "The lifting scheme: a custom-design construction of biorthogonal wavelets," *J. of Appl. and Comput. Harmonic Analysis*, vol. 3, no. 2, pp. 186–200, 1996.

[81] G. Deslauriers; S. Dubuc, "Symmetric iterative interpolation," *Constr. Approx.*, vol. 5, pp. 49–68, 1989.

[82] D. Donoho and I. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.

[83] Matthew A. Grayson, "The Heat Equation Shrinks Embedded Plane Curves to Round Points," *Journal of Differential Geometry*, vol. 26, pp. 285–314, 1987.

[84] Guillermo Sapiro Alfred Bruckstein and Doron Shaked, "Evolutions of planar polygons," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 9, no. 6, pp. 991–1013, 1995.

[85] Anothai Rattarangsi and Roland T. Chin, "Scale-based detection of corners of planar curves," *IEEE Trans. Pattern Analysis, and Machine Intelligence*, vol. 14, no. 4, pp. 430–449, 1992.

[86] Farzin Mokhtarian and Alan K. Mackworth, "A theory of multiscale, curvature-based shape representation for planar curves," *IEEE Trans. Pattern Analysis, and Machine Intelligence*, vol. 14, no. 8, pp. 789–805, 1992.

[87] A. El-Fallah and G. Ford, "The evolution of mean curvature in image processing," *Proc. IEEE Int. Conf. on Image Processing*, pp. 298–303, 1994.

[88] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Analysis, and Machine Intelligence*, vol. 8, pp. 679–698, 1986.

[89] D. Marr and E. Hildreth, "Theory of edge detection," *Proc. Roy. Soc. Lon.*, vol. 207, pp. 187–217, 1980.

[90] A. Rosenfeld and M. Thurston, "Edge and curve detection for visual scene analysis," *IEEE Trans. Computing*, vol. C-20, pp. 562–569, 1971.

[91] A. Witkin, "Scale space filtering," *Proc. 8th Int. Joint Conf. Art. Intell.(West Germanny)*, pp. 1019–1022, 1983.

[92] S. Mallat and W. L. Hwang, "Singularity detection and processing with wavelets," *IEEE Trans. Information Theory*, vol. 38, pp. 617–643, 1992.

[93] M. Cage and R. S. Hamilton, "The heat equation shrinking convex plane curves," *Journal of Differential Geometry*, vol. 23, pp. 69–96, 1986.

[94] M. H. Protter and H. F. Weinberger, *Maximum Principles in Differential Equations*, Springer-Verlag, Springer-Verlag New York Inc., 1984.

[95] P.J. Burt, "A gradient pyramid basis for pattern selective fusion," *Proc. of the Society for Information Display Conference*, 1992.

[96] Cakir H.J.; Khorram S.; Dai X.L.; de Fraipont P., "Merging spot xs and sar imagery using the wavelet transform method to improve classification accuracy," *IGARSS '99 Proceedings*, vol. 1, pp. 71–73, 1999.

[97] C.G. Small, *The Statistical Theory of Shape*, Springer, Springer, New York, USA, 1996.

[98] Horn A., "Doubly stochastic matrices and the diagonal of a rotation matrix," *American Journal of Mathematics*, vol. 76, pp. 620–630, 1954.

[99] Vincent L. and Soille P., "Watersheds in digital spaces: An efficient algorithm based on immersion simulations," *IEEE Trans. Pattern Analysis, and Machine Intelligence*, vol. 13, no. 6, pp. 583–598, 1991.

[100] Pavlidis T. and Liow Y., "Integrating region growing and edge detection," *IEEE Trans. Pattern Analysis, and Machine Intelligence*, vol. 12, no. 3, pp. 225–233, 1990.